DOCUMENT RESUME

ED 206 721 TM 810 629

AUTHOR	Ward, James G.; Gould, Jewell C.
TITLE	Plain Talk About Standardized Tests. Research Report.
INSTITUTION	American Federation of Teachers, Washington, D.C.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	Oct 80
GRANT	NIE-G-79-0041
NOTE	93p.
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Achievement Tests; Aptitude Tests; Criterion Referenced Tests; Elementary Secondary Education; Minimum Competency Testing; Norm Referenced Tests; Scoring Formulas; *Standardized Tests; *Test Interpretation; Test Reliability; *Test Selection; *Test Use; Test Validity

ABSTRACT

This handbook, in two parts, constitutes a manual prepared by the American Federation of Teachers, for improving teachers' use of standardized tests. Part I outlines basic concepts and issues surrounding standardized testing for teachers, parents and school administrators. The terms norm-referenced tests, criterion referenced tests, minimum competency tests, achievement and aptitude tests are defined and explained, then followed by a section regarding test selection, in which the aspects of test validity and reliability are introduced. The next chapter, concerned with test interpretation, discusses how scores and various types of derived scores commonly used to report test results, how they are derived, and cautions to be considered in their use. Applications of standardized tests to instructional planning, placement decisions, diagnosis of student needs, and the evaluation of instructional programs are also discussed. Finally, basic premises contributing to the proper use of tests are reviewed. Appendixes include lists of available tests, test publishers, and reference materials which review tests. Part 2 presents a hypothetical school district and two exercises in test selection, score analysis and presentation of same to interested parties. (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206721

TM 810 629 (1072)

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ✓ This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PLAIN TALK ABOUT STANDARDIZED TESTS

BY

JAMES G. WARD
DIRECTOR OF RESEARCH

JEWELL C. GOULD
ASSISTANT DIRECTOR OF RESEARCH

[illegible]

A REPORT OF THE
RESEARCH DEPARTMENT OF THE
AMERICAN FEDERATION OF TEACHERS, AFL-CIO
OCTOBER, 1980

PLAIN TALK
ABOUT STANDARDIZED TESTS

Prepared by the American Federation of Teachers
Department of Research

James G. Ward, Director
Jewell C. Gould, Assistant Director

NIE GRANT/NIE-G-79-0041

This study was prepared by the Research Department of the American Federation of Teachers under Grant Number NIE-G-79-0041 from the National Institute of Education, U.S. Department of Education. The opinions expressed in this study do not necessarily reflect the position or policy of the National Institute of Education or the U.S. Department of Education.

TABLE OF CONTENTS

Preface

I. Testing Today 1

II. Standardized Tests 4

III. Selecting Standardized Tests to Suit Your Purposes: Standards for Evaluation 17

IV. Interpreting Test Results 35

V. Application of Standardized Tests to Your Purposes 46

Bibliography

Appendix A-14

This manual was prepared by the Research Department of the American Federation of Teachers in collaboration with the Center for the Study of Evaluation (CSE). We gratefully acknowledge the expert assistance of Anne Goldblatt and Helen Nemorin in the preparation and typing of this manual.

opeiu#2aflcio

PREFACE

The proper use of tests is a topic of interest to all teachers. Tests are an important tool for teaching and learning and their appropriate use is critical to the educational process.

A number of times throughout the 1970's the Executive Council of the American Federation of Teachers adopted resolutions calling for more study of testing, more responsible use of tests, the improvement of testing processes, and the wider dissemination of information on tests.

In 1978, the AFT applied for and received a two year grant from the National Institute of Education to prepare training materials and conduct conferences on improving teachers' use of standardized tests. Part of the plan was to survey a representative sample of teachers to ascertain their preparation and knowledge in testing, their assessment of the importance of testing to their teaching, and their attitudes toward various issues in testing. The results of that survey were published in a report entitled, "Teachers and Testing: A Survey of Knowledge and Attitudes," by James G. Ward (Washington, D.C.: American Federation of Teachers, July 1980).

This handbook, Plain Talk About Standardized Tests, is the basic manual for the training conferences. It is intended to be a primer on the issues in standardized tests for teachers and others who need a basic knowledge of the topic. The material on which a number of the chapters are based was a technical draft prepared by the Center for the Study of Evaluation, University of California at Los Angeles. Substantial revisions were made in those chapters based on comments and critiques from previous conference participants and outside reviewers, and from decisions made by the authors.

This handbook represents one part of the commitment of the American Federation of Teachers to the professional growth of its members. Through AFT training conferences, QuEST conferences, and other educational services of the AFT, these materials will contribute to greater understanding of all tests and the better education of children.

Chapter I

TESTING TODAY

Testing occupies a central position in American education. In one form of another, tests are accepted by almost all involved in education as part of that process. It is part of the rational method in Western thought that one determines what one wants to achieve, one tries to achieve it, and then one has some process by which to measure whether it has been achieved and to what extent. Therefore, some kind of testing or assessment is necessary to complete the process. American education has predicated much of its practice on this model.

It is difficult to escape tests in American schools. Tests other than teacher prepared tests are increasing in number and visibility.

- Between 40 and 45 states conduct state assessment programs.
- Almost 40 states have adopted minimum competency testing programs.
- Over 90 percent of local school districts regularly administer standardized norm-referenced tests to their students.
- Many federal education programs require regular, formal testing of children for program entry, program exist, or for program evaluation.
- The public is demanding more accountability from public schools and see tests as a vehicle to determine program success or failure.

Testing is an integral part of schools and is likely to remain so in the foreseeable future.

Yet, testing is controversial. While everyday educators are using test results to improve instruction and educational decision-making, critics of testing decry the essential meaningless of test scores and accuse tests of destroying children. For example, one such attack on testing shows a photograph of a sad, little six year old girl, with a tear on her cheek, who has just taken a test. The accompanying test explains that this child has just scored below average on a standardized test and, thus, at the tender age of six has been destroyed educationally for life. The message of this one page advertisement is clear. Because this child has scored below average on one test, there is nothing teachers or schools could possibly do to help this child. Such anti-education diatribes miss completely the role of tests in teaching and learning.

It is sometimes forgotten that tests are tools and, as such, can be used or abused.

Tests are one of the tools of the professional teacher. From test scores inferences about students are made. To the extent that teachers understand the limits of the test instrument and apply the test results within those limits, tests are helpful. It is when those limits are not known, or heeded, that tests become abused.

Understanding tests and their limits is not an easy task. The understanding is within the ability of professionals to grasp, provided some training is provided and sensitivity to the process remains present. Test construction, properly done, is a complex and often technical process. This manual can help teachers become aware of the demands of the process without requiring the ability to produce a test using sophisticated psychometric techniques. It can also help in the selection of the appropriate test for the various purposes teachers are testing. A methodical procedure for selection is outlined, and will prove helpful to teachers while inspiring confidence in the later use of the test. This promotes efficiency in the long run, and helps eliminate charges of too much time for testing compared to the use of test results.

Eventual integration of test results into the teaching process can be encouraged to the benefit of students, teachers and those responsible for policy decisions at the district level. It does not need to be abused. Testing, like a student, needs an appropriate mix of understanding and discipline. It must be encouraged to attend to the task at hand if anything of value is to develop

Plain Talk About Standardized Tests is intended to help teachers, parents, school administrators, and others concerned with successful learning to better understand the area of educational testing by focusing on one particular kind of test--the standardized test. This testing primer will help you understand what standardized tests are, how to select appropriate standardized tests to suit particular testing purposes, how to evaluate tests which are required to be administered in schools, how to interpret and use test results.

The underlying belief in this handbook is that tests are one of the many sources of information which may be used to improve education, but that tests alone do not provide enough information for decision-making. Tests are important, but they should never be used alone.

The remaining chapters in this handbook explore the issues in the use of standardized tests. Topics include:

- An overview of what standardized tests are, how they are developed, and what their uses and limitations are.

- A review of what some of the important factors are in selecting suitable tests or in evaluating existing tests, and how to apply those factors to test selection.

- An explanation of commonly used standardized test scores, how scores are derived, how scores are interpreted, and what cautions should be considered in score interpretation.

- A discussion of using the results of standardized tests in instructional planning with suggestions and procedures for using test results to make placement decisions, to diagnose student needs, and to evaluate instructional programs.

- A review of the basic premises which contribute to the wise and proper use of tests, and which will help prevent their misuse.

Various surveys and research studies, including the 1979 AFT Survey of Teachers' Knowledge and Attitudes on Testing, have shown a wide variation among teachers of background in testing and knowledge about testing. Therefore, this handbook takes a comprehensive approach including both basic definitions and concepts and more advanced topics on testing. A more experienced reader may want to skim certain sections of the handbook and concentrate on selected sections. A person new to the subject will want to take a more deliberate journey through the text.

Chapter II

STANDARDIZED TESTS

Some Introductory Comments

Everyone has taken tests. Tests are so much a part of our common educational experience that we often fail to look at some of the basic and underlying principles of tests and testing. This chapter begins with some basic concepts and ideas and proceeds through an examination of standardized tests, including their development and characteristics.

What is a Test?

A test is a systematic means of observing and describing behavior. It usually consists of a presentation of a standard set of questions to be answered. The answers to the questions are measured against a standard and a numerical value is assigned which is a description of the observed behavior. This "score" is interpreted as a measure of a characteristic of the person taking the test.

There are probably as many varieties of tests as there are those doing the testing. Tests can range from a conscious observation of study skills during an in-class work period to a formal admissions examination for a specialized graduate school. Tests may be oral or written. They may be impromptu or involve years of development time. Almost every act a teacher completes to assess a student's performance could be called a test.

A test may be designed for individual or group use. It may measure achievement, aptitude, attitudes, personality traits, or psychomotor behavior. A test is one source of information for educational decision-making.

Why Test?

Tests can be used for a wide variety of reasons. Although some reasons are highlighted below, one could think of others that have been excluded. Test results can be used to:

- Diagnose academic and behavior strengths and weaknesses.
- Prescribe specific educational plans for individuals and groups of learners.
- Place students in special accelerated or remedial classes and programs.

- Determine student achievement.
- Evaluate program effectiveness.
- Select students possessing particular abilities and/or aptitudes.
- Certify student competence.
- Promote teacher and system accountability.
- Predict future student behavior.
- Inform students about the development of certain abilities.

It is important to know why one wants to test before one can approach such issues as test design, test content, format for reporting test results, and other more specialized questions.

Why Do Teachers Need To Know More About Standardized Tests?

There are many reasons why knowledge on standardized tests is becoming increasingly more important.

First, tests are an integral part of the education process. If tests are not properly used, time and money spent on tests are wasted. Good testing is critical for effective learning. If you have no way of assessing effectiveness of the teaching effort, then it is extremely difficult to improve that effort. Standardized tests are one kind of the tests that are almost universally used in schools to make those kinds of assessments.

Secondly, there is growing public concern about accountability in education. For example, teachers in some places are evaluated on the basis of their students' academic performance and mastery of established educational objectives. While this practice has many limitations, the teacher's role and responsibility in the education of each learner is being emphasized across the country. Some people feel that standardized tests can be one way to approach accountability. These tests, then, may have important consequences for teachers, students, and the educational system as a whole. While tests are not the only method of accountability we have, they do help to maintain educational standards.

Third, federal and state supported programs often require audits, monitoring, and final evaluation reports not only to determine the effectiveness of implemented subsidized programs, but to make financial decisions about the continuance or termination of federal and state aid for the future. The government and the taxpayer want to know how, where, and why our dollars are being spent in the educational arena to educate today's children. Again, standardized tests often have real consequences for teachers and programs in trying to answer such questions.

Also, there are increasing legal challenges to current methods of evaluation. Some of the more recent discussion involves the administration of standardized tests to minority children, particularly black and Hispanic children, the placement of disproportionate numbers of minority students in classes for the educably mentally retarded, the use of results from minimum competency examinations to determine whether a high school diploma should be granted, the perceived decline of basic skills at both the elementary and secondary levels, and identification and assessment of language proficiency for children of non-English backgrounds.

In various regions of the country, litigation and legislation involving these issues have lead to efforts to abolish or restrain the use of standardized tests. Genuinely intentioned educators, parents, and politicians are requesting a reexamination of the role of tests and testing results in education. Teachers are, and should be, very much involved in the consideration of these issues.

Standardized Tests: What Are They?

Standardized tests, like all tests, attempt to provide teachers with information about their students. What makes a standardized test different from a classroom test, for example, is that the test items are preselected, and administration and scoring procedures are prescribed, or standardized, for all students taking the test. Also, they usually provide information about how others who took the test performed. The purpose for standardization of content, administration, scoring and interpretation becomes immediately apparent when one considers that such tests are often used to gather descriptive or comparative information from large groups of students. When comparing students from more than one classroom, variability among test conditions seriously hampers the ability to say anything at all about the students tested. An example will make this point even clearer.

Suppose that a school district wishes to determine how well all of the fifth grade students in the district read. The district consists of a number of elementary schools. One school is very crowded and students attend class in portable trailers parked next to the baseball field. Another is located near the airport and instruction must stop each time a jumbo-jet lands. In a third school, fifth graders are assigned to carpeted rooms with computer terminals for computer assisted instruction. It may be easy to predict which students have the best chance of performing successfully on the reading achievement test. While educational opportunity cannot be totally equalized in this district, the differences in testing conditions among students can be minimized and each student's chance of doing well on the test can be maximized by using standardized content, procedures, and scoring rules. The subsequent differences among students can then be ascribed to learning environment, or to differences in achievement and ability, rather than to differences in test climate. In order to have the kind of information needed for instructional planning, as many factors as possible that interfere with our ability to attribute test performance to student learning must be eliminated. Comparability of scores and testing conditions is an important factor.

Standardized tests, then, provide for the sampling of behavior under a set of uniform procedures. There is a common set of questions that are administered with the same set of directions and time limitation to students, with a uniform scoring procedure.

Standardized tests, like most other tests, are produced in four basic steps. One source cites the following steps as central to the construction of a standardized test:

- Planning the test.
- Preparing the test items.
- Experimental tryout and revision.
- Administering the standardization edition.

The initial step in test construction is the specification of the content and skills the test will cover. Test developers usually consult with teachers, curriculum experts, and book publishers to determine what is being taught and what is important. Next they outline the topics and skills they plan to assess, as well as the number of items that will be used for each area. The developers will write detailed content specifications which are usually provided in the test manual. This provides the user with some basis for determining how well the test content fits particular curriculum goals.

When content has been specified, it is fixed for all forms of the standardized test. Because most standardized tests are designed to reach a broad market, the content and skills selected are representative of what is taught in a large number of school districts from geographically separate regions. For this reason, the content of standardized tests often may not match perfectly local curricular goals or individual classroom learning objectives. However, they do provide a good basis for comparing students from different schools or areas.

Test items are written to conform to the content specifications created. Test developers produce many potential items for each skill area, then test the items with groups of students, and finally select the "best" items according to some prespecified criteria. During the test development process many items are written, tested, revised and discarded before the final form of the test is complete.

Standardized test developers give their draft test items to groups of pupils selected to represent the population for whom the test intended. The purpose of this try-out is to identify and refine the test items for final test forms and to test the adequacy of test directions, time limits, and format.

From the item statistics and other information generated, test developers select the best items and construct final forms of the test. These final forms are tested once again with groups of students representing the population with whom the test is to be used. This final tryout is designed to gather information about the quality of the test itself rather than individual items. The actual equivalence of test forms in terms of producing comparable group performance is determined at this stage. Reliability and validity statistics are computed from tryout results. Group performance on the standardization tryout is often used to construct tables for interpreting test scores. Also the adequacy of administration and scoring procedures is checked out once again. Standardization tryout results are written up in the test technical manual and provide important information for test consumers about the quality of the test and the limitations on score interpretation.

While this brief description of how a standardized test might be developed is oversimplified, it does provide a idea of the complexity of the test development process. Central to this process are three fundamental questions which must be kept in mind whenever testing is under consideration. These three questions are:

- o Exactly why are you giving (developing) the test?
- o What type of information do you expect from the test?
- o How do you intend to use this information once you have it?

As you read through this manual, do not ever lose sight of these three questions.

Types of Standardized Tests

Standardized tests can be grouped into three classifications based on what they measure. The three general categories of standardized tests are:

- Aptitude tests.
- Achievement tests.
- Personality, attitude, and interest inventories.

Each one is distinguished by its purpose and the kind of information it gathers.

Aptitude tests attempt to assess students' abilities and potential. Their purpose is to predict future performance. While some skills measured by aptitude tests may be learned, others are developmental. Although aptitude tests are not as dependent on school learning as achievement tests, it would be impossible to construct an aptitude test that did not measure school learning to some extent. Aptitude tests are used for administrative decisions such as student selection, classification, and placement, for guidance purposes, and sometimes for evaluation of instruction.

Achievement tests measure a student's prior learning and developed abilities specific content areas. Their purpose is to assess how much a student has learned in school. Achievement tests are used in instructional evaluation, in guidance, and for administrative decisions such as student selection, classification, and placement.

Personality, attitude, and interest inventories are non-cognitive measures and are used primarily for guidance purposes.

Norm-and Criterion-Referenced Tests

Standardized tests, particularly standardized achievement tests, are often described as being norm-references or criterion-referenced. The distinction between these two is very important, although in practice it often becomes blurred.

The fundamental difference between these two types of tests rests with how the scores are reported and interpreted. Norm-referencing means reporting scores so that one can tell how a student's score compares to the scores of others (the norm group). Criterion-referencing means reporting scores so that one can tell how a student's score or performance compares to some specified standard of proficiency.

Norm-referenced tests were designed to make selection and classification decisions where it is important to assess one persons's performance in comparison to the performance of others. Criterion-referenced tests were designed to describe a person's mastery of specific content in relation to a selected standard.

It should be obvious that norm-referenced tests, since they are based on stated content objectives, provide information on mastery of specific content and that criterion-referenced tests can be used to compare student performances. Hence, the distinction between the two is not always as clear and distinct as the definitions imply.

Norm-Referenced Tests: Construction, Interpretation, and Limitation

As with all tests, norm-referenced test construction starts with the specification of test content. Broad objectives or content areas are selected and a large number of items are written to assess each. Items are tried out with a group of examinees selected to represent the kinds of students for whom the test is designed.

Data from this initial try-out is critical for developing a test that will array student scores in a normal distribution. Several item statistics are computed to achieve this end. One statistic, the item difficulty index, is used to delete some items that too many students pass or fail and to retain items that are answered correctly by about half of the sample group. Another statistic, the item discrimination index, is used to identify test items which do not discriminate well enough between high and low achievers. In order to locate problems in multiple choice items, the number of students answering specific response alternatives may also be inspected. Once appropriate items have been identified through the analysis of the item statistics, test form are constructed and alternative forms of the test tried out again with an appropriate sample of students.

On this second tryout, the test developer is interested especially in norming the test and obtaining estimates of test reliability and validity. The group performance of students participating in this norming tryout becomes the comparison group or norm group for interpreting test scores. For example, if three percent of the norm sample got four items correct on a 100 item test, the norm tables would say that a score of four would be interpreted as falling in the 3rd percentile. Because student performance on a norm-referenced test derives meaning principally by comparison to the norm group, it is critical that the norming sample be representative of the students for whom the test is designed. The ideal norm group might represent the student population on the following dimensions: geographic region of the country, school size, community size, student demographic characteristics, and school type. Because norms can become outdated, tests are usually renormed every three to five years. The interpretation of norm-reference scores emerges directly from the fact that the test contains a selection of items designed to provide a group performance which will be arranged into a normal distribution. Test scores provide information on how students compare to a national or local group of students who have taken the same tests, and are expressed as percentile, stanine, or standard scores which have been derived from the raw scores. Because test scores tell how well a student or group performance compares with the peer group standard, norm-referenced tests are suitable for making decisions such as:

- Who shall be selected into a program when there are only a limited number of spaces available?
- How do our students compare with students of similar age and background on this skill area?

There are some limitations on the use of norm-referenced scores that must be recognized. Because norm-referenced tests are designed to decide how well a student's performance in a subject area compares with other students, they may yield limited information about the amount of specific skills a student has mastered. Subscale scores are sometimes not sufficiently reliable, and often do not include a sufficient number of items to tell you if a student has mastered a particular objective.

Criterion-Referenced Tests: Construction, Interpretation, And Limitation

In criterion-referenced testing, as in norm-referenced testing, item construction procedures focus on selection of content for the test. Test construction begins with the specification of the concrete objectives to be tested. Ideally, detailed content specifications are designed to include all of the salient parameters of content or skill domain, such as eligible content, vocabulary, syntax, readibility, readibility, and how distractors are to be written. Based on the objective specification, teachers should be able to know exactly what skills and knowledges are being tested. Based on these detailed objectives or domains, descriptions of many items are then generated to reflect each objective. Like norm-referenced tests, these items are subsequently also tried out on appropriate samples of students. However, the purpose of the tryout is not to "weed out" items that do not discriminate between high and low scorer, but rather to identify items that appear to be measuring the same objectives and eliminate items that do not.

Items selected for criterion-referenced tests should be selected for their ability to represent the content rather than to separate students. The ultimate criterion for item selection is the relevance of the item to the objective being tested. Often items that might be discarded from a norm-referenced test because they are "too easy" or "too difficult" are retained in a criterion referenced test because they measure important skills.

Like norm-referenced tests, the best items are then assigned to test forms and administered to representative groups of students. On the basis of these results, test reliability and validity is determined. The results of students performance during this final try-out may also be reported in the test manual to give users some basis for comparing their students' scores with other students, although providing comparative data is usually not an area of emphasis. Hence, criterion-referenced tests often become "normed."

As a result of these test construction procedures criterion-referenced test scores tell how much of the content a student has mastered. Scores are often reported in terms of the number or percentage of items correct. Scores tell about how well students have mastered the basic concepts described in the test specifications. They usually do not tell you if this compares well or poorly with the performance of other students in the same grade, although some test developers do provide you with information about the performance of a comparison group. Criterion-referenced interpretations are especially appropriate for the following kinds of decisions:

- Have students mastered program objectives?
- What are individual students' strengths and weaknesses in this area?
- Which objectives am I teaching effectively and which not so well?

Criterion-referenced tests, too, suffer from some disadvantages. Because these are relatively new kinds of achievement measures, the technical quality of many of these tests is questionable. Technologies for determining test quality are still emerging. Although these tests purport to give you specific information about student performance, many tests do not contain a sufficient number of items per objective to be definitive. Further, test specifications often are not as fully described or as well-defined as they should be. In addition, test objectives are often narrowly focused with the result that a single test is not equally suited to the curricula of different schools or classes, making interschool or program comparisons difficult.

Some Observations

The uses of norm-referenced and criterion-referenced tests vary. Norm-referenced tests are generally used for selection decisions, classification decisions, guidance decisions and in aptitude, interest, and personality inventories where comparisons are important. Criterion-referenced tests are generally used for classification and guidance decisions also, for placement and certification decisions, and to assess achievement of specific content.

For instruction decisions a teacher probably will want to both assess mastery and discriminate among students in achievement, so both criterion-referenced and norm-referenced tests are used. Measurement specialist N. E. Gronlund makes the distinction between testing the minimum essentials employing mastery testing using criterion-referencing and testing for maximum development employing discrimination testing using norm-referencing.

Again, many tests are amenable to both norm-referencing and criterion-referencing and scores are reported both ways.

Minimum Competency Tests

Minimum competency tests for students are a special type of achievement test which have developed in recent years. Minimum competency tests are developed to yield criterion-referenced interpretations, since the purpose of giving a competency test is to find out what a student knows in relation to prespecified objectives. Test construction procedures should therefore be much the same as those for criterion-referenced tests, although sometimes test developers will select items based on their ability to discriminate between those who pass the test and those who fail. Score interpretations are also similar to those of criterion-referenced tests, except by definition, scores on minimum competency tests can be interpreted as acceptable or unacceptable performance, that is, passing the minimum standard or not. Developers of other criterion-referenced tests also sometimes indicate a passing score or cut-off score indicating mastery.

Minimum competency tests share the problems and limitations of other standardized tests because the consequences of test performance may be very severe. To students not passing a grade or not receiving a high school diploma, these problems are more serious.

The task of deciding what goal or objective any test should assess is always difficult. The task of defining objectives that a minimally competent person should attain is complex indeed. For example, what skills are necessary for survival? What skills are truly basic and necessary? For whom? For what? In the absence of empirical data these decisions represent values and opinions that are known to differ among different individuals and groups.

The objectives for the test must bear a relationship to what is actually taught. Courts have ruled on the need to test what is taught and to advise students in advance of the need to pass the test for promotion or diploma. Challenges to the testing requirement have been filed in cases where special groups of students, such as handicapped, have been provided individualized education plans that are substantially different from the regular curriculum. If a student meets the requirements of an individualized education plan, must the student also pass the minimum competency test to receive a diploma? Questions such as these will be with the competency movement for some time. The answers may well shape the future course of education. In spite of the rapid growth of competency tests since 1978, we are still closer to the beginning of this issue than we are to the end.

In addition to definitional problems, minimum competency tests, like other criterion-referenced tests suffer from problems inherent in any new technology. Reliability and validity are especially important, yet it is not always clear which statistics are best to describe these test properties.

Setting a passing or cut-off score for minimum competency tests is another problem area. Various methods have been advanced to deal with this problem, including the use of expert opinion, past performance of students on the test, and statistical models incorporating the probabilities of misclassification (i.e., identifying a "competent" student as incompetent and vice versa). Although consideration of all these factors will likely yield the most reasonable cut-score, current practice often deviates from this ideal.

In order to help insure student competency, the results of minimum competency tests should have implications for student remediation. As with criterion-referenced tests in general, the number of items assessing each objective often is not sufficient to determine reliably student mastery, and test objectives frequently are not defined well enough to suggest what specific instructional remedies are necessary.

Aptitude Tests and Achievement Tests

While achievement tests are designed to measure past learning of school specific content, aptitude tests are designed to predict future achievement, or the ability to acquire new content or skills. In theory, an aptitude test could be either norm-referenced or criterion-referenced. However, most standardized aptitude tests are constructed to yield norm-referenced interpretations, that is, how a student's potential compares with others. As a result, aptitude measures are often used to identify students with special needs and abilities, for placing students into special programs, and for guidance and counseling purposes.

The distinction between aptitude and achievement tests appears clear conceptually, but in reality their functions overlap. Although aptitude tests are apt to be less dependent on specific school content, both kinds of tests measure learning and previous experience, both in school and out of school. For example, while an achievement test may measure number of concepts and basic computational operations, an aptitude test might assess problem-solving ability, figure analogies and abstract reasoning. Depending on the curriculum of a particular school, these latter skills may represent aptitude or, if taught as part of the curriculum, achievement.

The distinction between aptitude tests and achievement tests in terms of predictive value also is somewhat muddled. Although aptitude measures purport to measure potential and predict future success, it is clear that achievement tests also can serve this purpose. That is, prior learning in a particular subject is often the single best predictor of future success in that area. For example, if a student in the second grade performed well on a reading achievement test, it would clearly be expected that the student would do well in the third grade reading program.

Although achievement and aptitude tests can serve similar functions, aptitude tests do have some advantages. For example, they can be used with students who have had no prior exposure to a subject, and they are often less time-consuming than achievement tests. Consequently, aptitude tests can offer an efficient method for screening and selecting students. However, because the content of these tests often is not tied to specific subject areas, the instructional implications of test results are limited. For example, if a student performs poorly on a mathematics achievement test, the results might indicate that instruction and practice were needed in geometry and ratio problems. If, however, an aptitude test shows a student is low in non-verbal ability, what instructional actions should be taken by a teacher?

Because the results of scholastic aptitude and mental ability tests can have serious consequences for students, they must be interpreted with great caution. If a student does poorly on an aptitude test, it may mean that the student has a poor chance of succeeding in certain school subjects. Equally likely, perhaps, poor performance could be due to the fact that the student did not have the environmental opportunities to develop the abilities in question. Here is raised the issue of cultural bias, a problem that exists in achievement tests as well. For example, one common aptitude test item shows three sailboats in a picture at different distances from the horizon, and asks the students to identify which boat is furthest. Reading of picture cues is a culturally embedded skill that is taught in school in come countries. A student who failed this item may have done so because the student was not taught how to judge perspective in two-dimensional space. Thus, it often is difficult to determine exactly how much of an aptitude test score is a function of cultural difference and prior chance to learn and how much a function of low ability. Cultural difference also may affect other aspects of test performance. For example, factors such as examiner-child rapport, anxiety, motivation, understanding of directions, etc., clearly will influence student test scores. Consequently, many have argued that mental aptitude tests are biased toward middle class culture, and discriminate seriously against minority group members, students from non-English speaking backgrounds and those from lower socioeconomic backgrounds. Some states have banned intelligence testing entirely, while others have mandated procedures to ensure that tests do not receive undue emphasis in decisions about student futures. Others have argued that those skills and knowledges that are needed to succeed on these tests are the same ones needed to succeed in school and society.

Chapter III

SELECTING STANDARDIZED TESTS TO SUIT YOUR PURPOSES: STANDARDS FOR EVALUATION

Test selection involves a consideration of the quality of the test instrument and the appropriateness of the test in fulfilling local purposes. The end result of the process of test selection is not so much a compromise of either standard, but rather an awareness of the limitations of any test so that accommodations can be made in the use of test results. It may be that the information gained by testing will not provide all that the district originally sought, but the appropriate use of the test has limits which cannot be exceeded if the interpretations are to remain accurate. Some judgments about tests must be made by persons qualified in the selection and use of tests. This chapter will identify the measures by which tests are judged and suggest some ways of selecting tests so that local goals can be more easily and more realistically reached. One of the most important considerations will be how closely the tests match the district's practices and goals. By knowing a little about the construction of tests a person can reasonably be expected to make an appropriate selection of a quality instrument. The key, however, is which instruments of comparable quality will perform the functions that allow the most economical and effective use of student and teacher time and effort.

Test selection is a little bit like sculpting an elephant out of large granite rock. With hammer and chisel you knock away everything that does not look like an elephant until the sculpture is complete. In selecting a test you will eliminate those tests with features inappropriate for your use of which do not add anything to the accomplishment of your goals. You will eliminate tests which do not measure up to the standards of test construction. In the end you must choose between tests which to some extent meet your criteria. It is judgment by professionals at this point which will produce something of value, something of continued usefulness, and something which will allow districts to proceed with the assessment of students to meet the identified needs.

Before one decides to test it would be useful to think about what one is attempting to accomplish. Measuring achievement or aptitude is not as easy as some might think. Imagine for a moment that you have come at the close of a late summer's evening to a small lake set in the woods. The half moon provides just enough light to see. An evening breeze blows a mist about the lake, obscuring but not completely blocking the shape of objects. You decide to recreate this scene and relate it to others. How would an artist begin to capture this scene? With limited tools and skills it is difficult to portray the changing shape of mist, the shrouded trees, and to add a sense of the other stimuli present such as the coolness of the moist night air against your skin.

The measurement of aptitude and achievement is just as elusive. We must proceed with measurement tools which are imperfect and observation skills which are limited. The artist would struggle to see in the mist the shapes that provide dimension and then try with brush or pen to portray what he saw. When observing the results of the artist, the viewer attempts to supply many of the intangibles of the scene based on his own past experience. In testing we observe the product of the existing tools of measurement and struggle with a desire to relate our observations to our own experience.

There is a demand for some proof of the existence of learning. Some have real or imagined uses for the outcomes which develop as a result of teaching and learning. It requires a skilled observer to assist in identifying the detail which would be missed by most persons. In this respect there is a role for teachers in testing and for critics in art. Both are sensitive to the medium and the subject and are perhaps the best resource to describe what occurs in either process. Such comments along with reflection based on our accumulation of experience allow us to infer certain things about what we could expect in future experiences of this kind. Supportive and reminiscent, sketches and interpretations of life known as tests help people think about their concerns. This access to events occurring in another place and time is at the heart of measurement, the reason that art exists.

If the decision is made that tests are to be used, then the first task is to establish the purpose of the testing program. Secondly, one must search for a useable testing instrument. The usefulness of a test will be determined by how consistently it provides the information sought and the degree to which the test is capable of achieving stated goals. These are the primary considerations in test selection.

Technical Properties of a Test

The properties of standardized tests which a school district might first consider in the selection of a test would include consideration of the reliability and the validity of the test. Estimates of reliability and validity are provided in information about the test supplied by test publishers and independent reviewers. Test publishers complying with the Standards for Psychological and Educational Testing will describe the results of several measures of reliability and validity along with other descriptive data in test manuals. Oscar Buros's Mental Measurement Yearbook lists and reviews a large number of currently published tests. Other sources of information are available and will be detailed later in this chapter. These sources may be useful in making comparisions about the qualities and properties of tests as they relate to each other and to the goals of the district testing program. This section discusses how these comparisons are developed and suggests some cautions to be considered.

Typically a district would use standardized achievement tests to make generalizations about a person's performance on measures of developed abilities. From personality inventories and aptitude tests mentioned in the previous chapter, other inferences are made. Any measure of behavioral or educational characteristics of a student--the non-physical measurements--are subject to some error. It is the nature of psychological and educational testing that the process is less exact than physical measurement. When using test results, teachers, counselors, and administrators need an awareness of the possibility for error. Test results are useful in drawing inferences and generalizations when used in combination with other factors to develop judgments. It is not correct to abandon the use of tests simply because they do not allow us to weigh in scales the thoughts of man. The weight of a gold ball should remain constant through several weighings. Unlike a gold ball, the educational and psychological characteristics of students are in flux. Our efforts to measure those characteristics will be affected by the changes that take place within each individual. Even if we could assess the same group repeatedly with the same instrument, which is not a likely or probable occurance, the results of the test would vary according to factors largely independent of the test instrument.

PART A

RELIABILITY

Reliability is used in testing to indicate how consistently a test can measure performance over time and when administered in somewhat different conditions. When the reliability of a test is high, individuals will retain their relative rank when scores are compared to others in repeated administrations of the test. A test must provide consistent results even considering that students will often take the test in various locations and with different test administrators. These differences will produce slight variations in instructions, timing, and other environmental factors. To be useful to a district, a test must be tolerant of such variations and produce useable results in spite of the differences. Some of these variables affecting the consistency of tests are trait instability, sampling error, administration error, and scoring errors.

Trait Instability

Trait instability is another way of saying how much educational characteristics will vary over time. It is one of the variables in testing that is independent of the test. What a student knows and forgets about a body of knowledge or domain will change with the experiences he has. The reaction to the different questions used in a test will vary.

Sampling Error

Sampling error is the term used to describe how the choice of questions affects the reliability of the test. Since test questions will produce scores from which inference about the knowledge of a specific student or groups of students is made, then some provision for identifying this factor must be included.

Administrative Errors

Errors in the administration of a test may be made. While such error should be limited, the complete elimination of error in this area is not likely. The test must be flexible enough to account for a range of individual styles and conditions in test administration.

Scoring Error

Scoring error is a mistake on tests when a student knows the correct answer, but marks the answer incorrectly. A less than accurate picture of how well the student knows the information being tested is presented. Such errors should not be overlooked as to their effect on test reliability.

Error Variance

In a reliable test, finally, the personal qualities that the student brings into the testing session should not drastically affect the outcome. How a person feels physically, how interested he is in taking the test, even how lucky he is in guessing the correct answer will all affect the reliability of the test. In the construction of a test, the designer must not only take into account these non-test variables, or error variances, but must also explain in the test information the result of considering these error variances and how reliable the test is after taking into account these possibilities.

Reporting Reliability

Many kinds of variables will affect test scores. It is reasonable to expect that test publishers will provide information they feel is important about the effect of variables on the reliability of their test. Since reliability can be established in a number of ways that will be different for different kinds of tests, the district's purposes, and not those of the test publisher, should be the criteria against which relizbility is judged.

For example, if you were looking for a competency test you would likely select a test that has used equivalent forms to establish reliability. You want to have a test that is very reliable around the passing point as opposed to the high or low end of the scores. Equivalent forms of the test tend to provide a superior measure of the test's consistency around the passing point or central point of the scoring range. In a test that is to measure mastery of a skill or domain, it might be more useful to determine very accurately the scores at the top end of the range. The various methods of establishing test reliability are presented so that in selecting a test the method that best fits the district's needs can be examined.

Measures of Reliability

Reliability information, as reported in reviews of tests, will frequently be reported in quantitative terms derived through statistical analysis of the results of test administrations. The results of this analysis, commonly reported as a correlation coefficient, will show how closely the results compared agree with each other. The correlation coefficient is a measure of relationship. What the test reviews will hope to show is that a reliable test will produce test scores that have a high correlation coefficient. The correlation coefficient is expressed as a value ranging from +1.00 to -1.00. To indicate performance on the test which is identical for the same person or groups of persons, a +1.00 correlation would be provided. To show complete opposite performance, a correlation of -1.00 would be assigned. To show that there was no relationship at all a correlation of 0.00 would result. For a reliable test instrument, a strong positive correlation is desirable. As a general rule, a correlation of +0.85 would be acceptable when considering a test for which group inferences will be made. The specific needs of the district will govern what correlation is acceptable, but a correlation close to zero will indicate a test which will be less than useful.

Tests of Reliability

To assist with test-to-test comparison, publishers provide much of the data one needs in assessing reliability. The reliability indices reported in the test manual or by reviewers of tests are likely to be determined by one of the following methods: test-retest, equivalent forms, split halves, or Kuder-Richardson. They measure equivalence or stability in the test and would be reported as appropriate to the test.

Test-Retest Method

Test-retest reliability is determined by administering a test to a group of persons, and then re-administer the same test to the same group at a later time. The scores for both tests for the group are correlated. When reporting reliability in this manner, the test should indicate how long an interval of time will affect the way a person is likely to answer a particular question. If the interval is too short, the learning that occurs in the first test administration is likely to be carried over to the second test. The student may mark an answer incorrectly on the second test the same way as on the first test because of memory effect. If the test is to measure developed abilities, this kind of carry over will give results that do not reflect the students' true ability. Tests in the psychological domain are not so severely affected by this kind of error.

In addition to the score variance that occurs as the same person retakes the same test over time,other possibilities for error variance present themselves. Different questions on separate forms of the test might be more difficult for some students. To correct for this error, measures of equivalence have been developed in an attempt to identify such potential problems. Equivalent forms of the same test which are equal in content and statistical properties are administered to the same group of students. Sometimes such forms are described as being parallel to emphasize that the content is to be similar between the two forms. An example of this would be to weigh the same object on two different scales on the same day. Assuming the object does not physically change between the two weighings, we would expect the differences in weight to be due to the difference in the instrument used for measurement, or to measurement errors.

Equivalent Forms Method

In parallel or equivalent form measures of reliability the student will take both tests and the scores will be compared. If there is a high correlation we assume the test is reliable. In selecting the two forms to be used, consideration should be given to similarity of content, mean scores, and variances for each. If the tests follow closely in succession, it is likely that the differences in scores will be due to the differences in the forms. The range of difficulty of items, format, time for administration, and examples must all be carefully considered in the construction of the tests to be compared. Because of the difficulty in creating such equivalent forms, test designers have sought other methods to establish reliability.

Split Half Method

When it is impractical to test the same group on more than one occasion, or if alternate forms of the test are not available, items making up the test scores can be examined to determine reliability. Student performance on individual items on the test is related to the total test score. Split-half reliability is one such method of establishing the internal consistency and the homogeneity of test items. The scores of the items are separated, the sub-scores determined, and then correlated. If the sub-scores are identical, then we have a measure of how reliable a test only half as long as the original might be.

Kuder-Richardson Method

To avoid splitting the test to check reliability, the Kuder-Richardson formulas can be used when the items are scored either "right or wrong," or in some "all or none" type tests. The Kuder-Richardson formulas, K-R 20 and K-R 21, are used in situations where items on the test are assumed to have either the same level of difficulty or different levels of difficulty. K-R 21 is the formula which assumes constant difficulty levels for all items. It is more useful for classroom teachers to know because the computation is direct and easy. The teacher must only compute the mean and variance for the test and substitute the values determined and the number of test items into the formula.

The K-R formulas produce correlation coefficients comparable to procedures used earlier. Because of the nature of the K-R formulas, there will be generally lower correlations from tests that measure widely varying skills and content than from tests which require the same kinds of skills and cover similar content. For example, a test which consists entirely of vocabulary words will likely produce more consistent results internally than a test that includes vocabulary, spelling, and math computations. The first kind of test is said to be homogeneous and the second is said to be heterogeneous. Individual variations are going to be greater on heterogeneous tests because all skills of an individual do not develop at the same rate and to the same extent. Even though reliable assumptions can be made from test performance on tests that cover a single skill, the purpose for testing in your school may be to assess a variety of skills and aptitudes. In such cases a lower, but acceptable correlation measure may be the best choice.

Comparing Reliability Methods

The purpose for testing as well as what the test is supposed to measure must be considered in examining the measure of reliability and item consistency. If you wish to determine how heterogeneous the test is, the difference between the correlation for the split-half reliability and the Kuder-Richardson correlation would be an indication. Kuder-Richardson formulas will show a lower correlation for widely varying skills. Split-half measures of reliability would likely produce a higher correlation on such rests. Knowing the difference can help to identify the heterogeneity of the test under consideration.

Other Factors Affecting Reliability

In a longer test, a split-half reliability measure will produce a higher reliability coefficient or correlation than when the separate parts are considered. This will continue to hold true as the test is lengthened by the addition of equivalent items. It will work in the reverse as well. When equivalent items are taken out of the test, the correlation coefficient will decline, Practical considerations of how much time can be alloted to testing will determine the most appropriate test length for individual school purposes. It may be that a highly reliable test which is too long will be passed over for a shorter, but less reliable alternative.

The speed quality of the test will also affect the reliability. As the internal consistency of a test is examined, some differences in item difficulty will be found. A test which emphasizes speed over power (knowledge brought to the testing situation) will be designed in such a way as to allow most students to answer all of the questions. In a speed test, difficulty level of items is quite low and the items are very consistent. Because of this the reliability coefficient will appear to be quite high when items of the test are compared to one another. If a test is selected for its speed qualities alone, [illegible]eliability measure such as test-retest or alternate forms would tell you more about the test's ability to accomplish its goals than one which simply examined halves of the test or looked to other internal consistency measures.

Another factor affecting reliability measures is the homogeneity of the group. If you have a group of students that is quite similar to the group on which the measures were established, then in a reliable test the results can be expected to be consistent. For example, suppose an achievement test under consideration is to be administered to a group of seventh graders. One would expect individual seventh graders' scores on the test to vary more randomly than if the test were given to students of markedly different ability levels such as might be encountered in a grade span of fourth through ninth grades. Assuming the items to be consistently difficult, the fourth graders should score lower on the various forms of the test and the more advanced students higher. When testing a group similar in age and abilities, such as students in a single grade, the probability of consistent scoring differences occurring drops. The characteristics of the norming group can be compared to the groups to be tested to determine if the reliability of the test can be expected to remain consistent.

Similar to group homogeneity, item homogeneity will affect the reliability of the scores. The difficulty of the items used will determine high or low reliability based on the number of persons who correctly answer the question. Consistency of scoring will affect the measure of reliability if everyone answers the majority of items because they are easy, or everyone misses the majority of answers because they are more difficult. The test examiner is hampered in determining reliability if student ranking is not apparent from the test scores. This is because reliability, by the definition we are using, calls for the same individual to retain the same rank in relationship to other scores through repeated admissions of the test.

These variables are taken into account in the development of the test. The results of the review of the test show the degree of consistency of results over repeated admissions of the test. The selection of the test most appropriate to your purpose can be made using this information. Once sufficient reliability is established, other measures of appropriateness of the test can be considered.

PART B

VALIDITY

A test may be reliable and can provide consistent results, but will be oflittle use if it does not measure what we are interested in measuring. If it purports to measure what we hope to learn about, but is inaccurate, then it is still not useful.. No test can be said to be absolutely reliable or valid in the abstract. It is quite possible to have a test that is highly valid for a particular purpose, but invalid for others. Validity is described as the degree to which a test measures what it intends to measure. There are four kinds of validity measures. They are content validity, predictive validity, criterion-related validity, concurrent validity and construct validity. There is also a measure or consideration known as face validity, which is technically not a validity measure, but is related. Predictive validity and concurrent validity are often grouped together and called criterion-related validity.

Content Validity

Content validity is the extent to which a test measures the subject matter content and the behavioral changes under consideration.

Content validity is of particular concern in achievement testing. A test is carefully analyzed to determine the subject matter content covered and the responses test takers are expected to make, compared to the domain of achievement to be measured.

To judge content validity, first the content domain must be defined. This involved consideration of both the subject matter and the type of behavior or task to be measured. Both the content and the process are important. Content domains and behaviors which are tested by the instrument should be identified. It is the responsibility of the group involved in the selection of the test to compare the domains and behaviors tested with their own goals to determine the best available match. Because of the practical limitations on test length, only samples of these domains and behaviors will be included in a test. Test specification tables and grids which identify the objectives of the test and the content and behavior which will be tested can be used to decide if a sufficiently large and representative sample has been included. How many subdivisions are included for each major category will vary. Because a test for a district will be given to students from a large number of teachers, the detail of the content and behavior to be considered will be different for groups of students. It would be helpful to the classroom teacher to see the widest possible consideration of specifications so that the practice for the greatest number of classrooms can be matched to the specifications of the test instrument. Then, after the inspection and comparison, judgments can be made about content validity.

In addition to seeking the individualizations which occur from classroom to classroom, it is important to remind those involved in the selection process of the need for some continuing standard against which the test can be compared. The text for the classroom, or the supplementary materials provided by the district will give some clues as to what the actual content taught in the courses might be. Dr. Andrew Porter of the Center for Research on Teaching, Michigan State University, has been working to develop methods of identifying the relationship of the major reading textbook series to the most commonly used achievement tests. By using a matrix of items from chapter exercises and the test items, comparisons are derived. Such kinds of studies may help to more clearly describe what items from the texts are sampled by the tests. The process appears to have application to other non-textbook materials and curriculum inclusions as well. If successful, a major obstacle would be eliminated and test content could be more accurately matched to the classroom instruction.

Examining for Content Validity

Four possible threats to content validity should be taken into account when examining a test. What is the extent of the mismatch between program objectives and the objectives of the test? While it would be unusual to find a test that matches perfectly with the program goals of the district, the test should address itself to district goals as closely as possible if program evaluation is the objective.

Does the instrument really test the skills that it is intended to? Is it much broader or narrower in scope than it claims to be? Suppose writing skills are to be tested. Upon review of a test you judge the skill tested most frequently bo be that of proofreading. The test would be more valid for assessing editing and correcting skills than for assessing skill in sentence construction, style, etc.

Is the vocabulary or the format of the test familiar to students or does the test rely on a specific set of curriculum materials? In some instances tests will be developed to match specific materials sets, and would include vocabulary that would be unfamiliar to the students. Content and format of the test may be equally dependent on materials used, so a thorough examination of the entire test is in order.

Are there enough items for each objective to be tested accurately? In some cases a test may be generally valid for all other purposes of the district, but may include too few items relating to some subskills. If less than five to eight items are included, it may be difficult to make some statement about a student's skills in that area. For inferences about groups of students, a few less might give some indication of how the group might do, but the emphasis of the test should ideally match the emphasis of the district's program as often as possible.

Predictive and Concurrent Validity

The next two forms of validity are grouped under the heading of criterion-related validity. When using some criteria to validate the test, it is possible to collect the data at the same time, or concurrently. This procedure is used to determine if the test provides the same information as some other measure, and the extent of agreement between the two. Predictive validity is generally determined by collecting the data at two different times. It is often used to see how well the test can predict future performance. Concurrent validity determines to what extent the information from one test can be substituted for another. Predictive validity seeks to establish if performance on a test can predict some characteristic.

One problem with the use of criterion measures for comparison is the lack of agreement that will surround the various possible choices. For example, not everyone will agree that grades at the end of a course are fair criteria. They might argue that such a criterion could be affected by the subjective judgment of the teacher. Suppose teachers had access to previous grades or test scores. An inclination would be present to grade according to how well they thought a student would probably do, and not on how well the student actually performed. Others might argue just as persuasively for the inclusion of grades and cite the relationship between grades and future success in school. If there is a disagreement, then persons determined to be expert in the field might be consulted as to what criteria would be consistent enough to use.

Construct Validity

The kinds of validity measures presented so far have been all related to some specific practical use of tests. Another type of validity has application to test interpretation in relationship to some psychological theory. Construct validity attempts to explain some psycholofical quality which we feel exists in order to explain some aspect of behavior. In establishing if a test has construct validity, one first needs to determine what constructs might account for certain behavior on a test. Some assumptions or hypotheses can be developed for that construct. Finally one would seek to verify the assumptions by logic and by empirical procedures. In the process both the test and theory are validated.

One assumption about intelligence is that it will increase with age. Another is that test scores on some standardized tests will differ with certain groups such as the educationally handicapped and the educationally gifted. Other characteristics of intelligence might be identified as well. The results of tests given to the different groups can be examined in light of the assumed characteristics of intelligence. If the results match or have some high degree of correlation, then the test might be considered valid. If they do not match, the test might be considered invalid and the theory correct. Another assumption would be that the underlying theory of the psychological construct is wrong. Whichever one chooses to believe, the independent judgment of the teacher and of groups of teachers must be applied to the question, "Precisely what does this test measure?" A review of the total evidence available about the construct considered should give some clue as to the usefulness of this measure of validity.

Face Validity

Finally the test should appear to be valid to the casual observer, to the student who takes the test, and school personnel who participate in the administration and use of the test and its results. While face validity is not technically validity at all, but rather a judgment about how the test might be considered, it is nevertheless important. If parents feel the test is irrelevant, then they might not seriously consider the outcomes. Students taking the test might not perform with enthusiasm and concern if they feel somehow the test is not a serious effort on the part of the district to gain information. The appearance or face value of the test is considered, and the test selectors must decide if by outward appearances persons will seriously consider the test and the results obtained by using it.

PART C

TEST NORMS

Test norms are provided by the test publisher to allow those involved in selecting tests to determine if the group upon which the test was normed or standardized is similar to the group of students in their district. Not every instructor's manual for a test will describe in detail the specifics of the norming population. This unfortunate occurance has been allowed to become practice in too many instances and accounts for a great deal of the confusion over lower than anticipated test scores. What if, for example, a test was normed on a sample that was geographically removed from the district considering the test. Certain speech patterns and idiosyncrasies can be introduced into the vocabulary of the test which would be advantageous to students from the norming area. You need go no further than the definition for a flavored, usually colored, carbonated beverage popular with both children and adults. Students from the east coast of the United States would certainly mark "soda" as the correct answer. The boys and girls from Kansas City would be surprised if the correct answer were anything but "pop."

The influence of television advertising on children might affect the answers given by different age groups. Early primary children would possibly believe that "S-E-I-T-Z" spells baloney. After all, they were told that in certain advertisements by Seitz, a manufacturer of processed meats. Older test takers might be inclined to spell "relief" "ROLAIDS," especially after completing the test section on mathematical computation.

Some general rules to keep in mind when considering the norming population in regard to the district population would include sufficient size, diversity, age, geographic location, and other characteristics to draw comparisons to the district which will be giving the test. Such information is available from test publishers who do not provide local norms, but can offer assistance in helping the district to determine them. The district can have its test scores reported in such a way as to review the entire district and make some assumptions about how the local population can be expected to perform over time. It is a process of matching local and national characteristics. More importantly, a judgment must be made as to what the effect of a mismatch will be when matching national information to the local population.

Publishers will usually indicate at what time of year the test was normed. Some will provide information for more than one time; for example, the Metropolitan Test reports spring and fall norms so that the differences which occur over the school year will be accounted for in the comparisons. Tests that have been recently normed will be the most useful. The content of the test can be examined. If the content is severely dated, further investigation of the norming date should be made. The copyright date on a test is not always an indication of the norming date, so specific information should be obtained from the publisher. What constitutes an adequate size is relative to your needs, but a good rule is to look for a test that has been normed on a large sample taken from various locations. A test given to a large number of students at a very few locations is not as likely to be as representative as one which was given to a group of comparable size at a variety of locations. The use of a few schools might severely skew the data and make comparison impossible.

Norms will tell how persons as a group performed on the test, and not how they should have performed on the test. They will allow comparisons to be made at the district level about performance, but will not indicate what the level of performance should be. They will not indicate advancement without looking at previous scores. It might be more useful for the teacher to know a child improved two grade levels than to know that the child is scoring at the average for all students nationally.

The choice of norms, either national, special group, or local, for reporting and comparison of district scores must be weighed carefully against the goals of the district. National norms, however large and diverse, will likely have been collected through tests administered at schools. Considering that different age groups attend or drop out of school differently, one realizes the possibility for excluding some people exists. If the national norm was established for eleventh graders and the national dropout rate was ten per cent, then the dropout rate in the district using the test might be a consideration if it exceeds or lags behind the average. Special group or fixed reference norms might provide a consistent standard if they fit the district's particular needs. Local norms might be fine for decisions about the school program from location to location, but might not satisfy the demand of legislators who seek information about how well the students in the state or district are performing compared to the national average.

PART D

STANDARDS FOR EDUCATION

It seems appropriate to make a point at this time about standards. One of the advantages of standardized tests is that their use over the years have produced results that are generally recognized by the public and many educators as acceptable standards. If there are misgivings about the standards demonstrated by test scores, it is a matter of degree and certainly not an absolute denial of their worth. Experts in educational measurement can point to significant improvements in tests and test use. These changes came about for a variety of reasons. Some changes have resulted in fewer deficiencies in tests, but most of the attention lately has been centered around the use of test results. This is where the greatest potential for abuse lies.

Repeatedly in this manual it is stated that the selection and use of tests must fit local goals and objectives. The inclusion of teachers and school officials in the selection of tests that best fit their goals and curriculum is stressed. This remains true and vital if tests are to be properly used and of value to a district program. If, at the same time, after extensive examination of available tests there are apparently none which are reasonably close to the content of the curriculum taught or seemingly none of the available instruments will fit the needs of the testing program, then perhaps an examination of the district curriculum might be in order. A great many innovations in curriculum and programs have developed. Not all are appropriate for the needs of students who must compete in a real world upon graduation. While it is not likely that any test will exactly match the specifics of each district, it is likely that for the test to be marketed profitably and economically offered that the test must have broad appeal. Suppose the district needs and goals are so divergent from other districts that none of the available tests can provide enough information about your program to be useful. It is possible that students are being prepared for a society vastly different from the one in which they will have to participate. This kind of injustice far exceeds any abuse that an inadequate test or the improper use of test results might generate. Concensus of many as to the needs of the group is often the vehicle which offers the best possibility for compromise. Students will be the beneficiaries if teachers are widely involved in the decisions affecting curriculum offerings, testing programs, and the overall goals of the system. Fears about teacher groups abusing curriculum goals are unfounded. Natural limits on the power of any single constituency exist in free collective bargaining. Teachers should pursue bargaining beyond economic issues and press for participation in the discussions relating to educational matters. Through this forum the observations of the classroom instructor can be presented.

PART E

PUTTING IT ALL TOGETHER

The selection of the test or tests that best serve purposes determined by the district is dependent upon a wide range of variables. Systematic procedures for consideration of these variables will certainly help the process, especially if established at the local level to meet the requirements and the resources of the local district. For this reason only a concept is offered as a basis for proceeding. No detailed formula or recipe offered here is going to contribute much to the establishment of a specific plan. Education about tests and involvement of the widest range of opinion and participants in the selection of the test offers the best hope to accomplish the task. The rough elements of test selection should include the participation of a group of individuals who are knowledgeable about the advantages and limitations of tests. If the knowledge is insufficient, then education must precede action. Careful review of the district's goals in education, and the part that testing plays would be a logical step to follow. An examination of the available tests can be made, but needs to be done considering the actual curriculum covered. A sequence emerges from this kind of reasoning, and sequence is probably the key to test selection. Education precedes the collection of information. If enough information has been gathered to make a decision, then proceed to the next step. If the knowledge about the application of the information seems too limited, then logic calls for more information or education. Teachers should not feel as if they are being rushed to judgment by either district officials or by others with an interest in the sale or acquisition of tests. Reputable test publishers can be expected to provide assistance and reference to assist in the selection of the instrument. Those who are reluctant to be of help must receive the message that without the support and assistance necessary to make a decision with which teachers will be comfortable, there will be no selection.

The technical considerations of tests are very important and should carry much weight in any final decision. There are some practical considerations as well. This final section discusses cost, format, time requirements, and similar concerns that will have an impact on the way tests are used and the extent to which they will prevent or encourage teachers to use test results in decisions about children.

The cost of tests must be considered in relationship to the cost of educating a student. Tests may range from as little as $.50 per student per administration and scoring, up to and beyond $1.35. When one adds the cost for sufficient manuals, individual tests, scoring and other associated costs, such as the proverbial #2 pencil, it may seem like a substantial sum in order to get the information sought. While it is generally wise to be conscious of the total cost for test administration, much more money is spent each year on the education of students. If every student had a test each school year that cost around four dollars, less than three tenths of one percent of his education cost would have been spent for testing. Information from this expenditure can be gained about the student that is used to make decisions about the most useful and productive way to allocate the other 99.7 percent of the resources available. The costs of tests are listed in several publications and the publishers will be able to provide up-to-date quotes on the current prices. Buros' Mental Measurement Yearbook lists the costs for tests, specimen sets, technical manuals, scoring charges, and incidental costs for tests listed. A comparison of costs in books which have been published for some time should give you an idea of how the various tests costs range, but the most current prices from the publishers along with anticipated future cost increases would be a more accurate estimate as the selection process moves toward a decision.

In the appendix is a list of some of the more commonly used tests along with the publishers. For a minimal cost, sample tests and technical manuals should be available. This kind of collection in the district's test resource library should be accessible to teachers. The district teacher center would be an ideal choice for such a library.

In addition, a service offered by the Educational Testing Service of Princeton, New Jersey, called simply the "Test Collection," contains an extensive library of tests and other measurement devices. It was established as an archive for testing and has current test information and related services available to persons engaged in education, research, and advisory activities. Over 10,000 tests are kept there in addition to files on American and foreign test publishers. Scoring services and systems, state testing programs, published test reviews, and reference materials on measurement and evaluation information are also available. These tests and materials are available to teachers and to districts and could serve to educate staff as well as expedite the test selection process. The staff of the Test Collection are available to answer phone and mail inquiries. Access to the Test Collection resources is also possible on-site to qualified persons who have an interest in testing.

Current information on testing can be obtained through a publication by the Test Collection called News on Tests. Announcements of new tests by publishers or non-commercial sources, citations of test reviews, and new reference materials of interest to those involved in testing are included in the annual ten-issue publication. Tests which are not commercially available, but cited in educational and psychological literature, are available on microfiche from the collection in individual copies or sets of a hundred. The list of Major U.S. Publishers of Standardized Tests is also available in pamphlet form. Annotated tests bibliographies in specific subject areas have been prepared and are available on request.

Also located at the ETS headquarters is the ERIC Clearinghouse on Tests, Measurement, and Evaluation. Annotated bibliographies are available from the ERIC Document Reproduction Service (EDRS), Computer Microfilm Corporation, P. O. Box 190, Arlington, Virginia 22210. These bibliographies would serve to provide the latest information on publications relating to the specific topic of testing that is important to the district. Some are included in the appendix to this manual.

In reviewing tests for selection it would be useful to speculate on the potential use of the test and what can be done to get maximum information from each test available. The length of time necessary to administer the test will play a role in how often a test might be used. If the time required to administer, score, and get the results back from tests seems excessive to school personnel, the tests may not be used as planned. Likewise, if the teachers who are administering the test feel excessive time is spent in the testing process that could be used more successfully in other educational activity, they might resist using the test. Some consideration for these attitudes must be made prior to the acquisition of the test. If the specifications of the test exceed what you determine will be tolerated by staff and students you should reconsider your choice. An alternative test might be adopted or the advantages that can be gained by using the particular test must be convincingly explained.

Test scoring is another practical consideration that must not be overlooked in selection. A variety of scoring services are available. Some simply provide a self-scoring guide for teachers. Others have elaborate computer-scoring and data comparisons. Choose the one that meets the needs of groups and individuals in the district who intend to use test results in making decisions. Test publishers should be willing to describe the various appropriate applications of the scoring services and specify the costs associated with each. Generally, a summary of the scores for pupils at school and district levels can assist in making some observations about the level of achievement in the district as well as establishing some expectations for performance of the various groups. The kinds of scores generally available include raw scores, national and local percentile scores, national and local stanine scores, standard scores, and grade or age equivalent scores. In some cases such as criterion-referenced tests a percent correct score might be provided. The reporting options are available for school, classroom or the individual needs. Some might be appropriate for parents or others concerned with the education process. The important thing to remember is that not all score reports are equally comprehensible and useable. Knowledge of the format for presentation and the information provided can help to promote wider use of the test for the purposes intended.

As previously stated, a systematic, thoughtful examination of the available options is going to produce for each district different procedures, but useful results. It would be of some value for the participants involved and for future test selection committees to have notes and summaries of the selection activities. Locally developed procedures such as these could be reviewed by experts and their advice sought as to possible improvements.

The end result of the effort should be the selection of an acceptable test that will provide an estimate. Tests are not unique from assessment in general, as all assessment provides estimates of the measure considered. Considering the problems that would occur without standards for comparison, tests properly constructed and used are far and away better than the absence of any standards.

Chapter IV

INTERPRETING TEST RESULTS

Once a test has been administered and some kind of score obtained, some sort of interpretation of that score needs to be made. It is difficult to place a test score in its proper perspective without some standard or basis for comparison.

Consider the following scores received by a student on an achievement test and its sub-tests:

Language Arts	Reading Comprehension	Mathematical Concepts	Mathematical Computations	Composite Score
52	46	61	48	53

These scores, by themselves, tell us nothing about the students' performance. We have no idea how many questions the student answered correctly out of the total possible. We do not know how much mastery over the subjects this student has. We do not know how this student's performance compares to that of other students. We do not know how this student did on one sub-test relative to another because we do not know what test scales are used or what the scores mean.

We must have some system for reporting test scores that will provide us with useful information.

This chapter discusses raw scores and various kinds of derived scores which are used to report test results. The intent of this chapter is not to provide a technical discussion of such scores, but to offer an overview of the concepts behind the various scores to help the reader in score interpretation.

Raw Scores and Derived Scores

The raw score is simply the number of test items a student answers correctly on the sub-test or test as a whole. For example, the student who correctly answered 37 of 88 items would have a raw score of 37.

Raw scores alone have little meaning. For example, what does it mean that a student achieved a raw score of 37 on a test? In order to interpret this score, you need to compare it to some standard. Derived scores, scores that are dervied from the raw score, provide some comparative information. They tell you what a student's raw score means in relation to the scores of other students, or what it means in relation to a student's accomplishment of test content. These two basic comparisons represent two divergent perspectives in testing that were described in Chapter II.

From a criterion-referenced perspective, derived scores would show to what extent a student has mastered a specific area of content. A percentage correct score would indicate what proportions of the content domain the student has mastered. Using the example above, the student would have a percentage correct score of 42 per cent:

$$\frac{37}{88} = 0.4205 \times 100 = 42.05 \text{ or } 42 \text{ per cent}$$

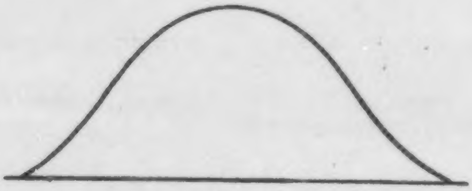
Also, we might want to have a minimally acceptable score, or cut-off score, to indicate the minimal mastery level that would be accepted.

Using the test above, a raw score of 50 might be the minimally acceptable score. The student attaining a 37 would fall below that score and would not have demonstrated the sought after level of mastery.

There are also techniques for comparing a test score of one student with others in a single group of test scores or with a larger group who have previously taken the test. Examples of such scores are percentiles, standard scores, stanines, normal curve equivalents, and grade level equivalents. These will be discussed below.

The Normal Distribution

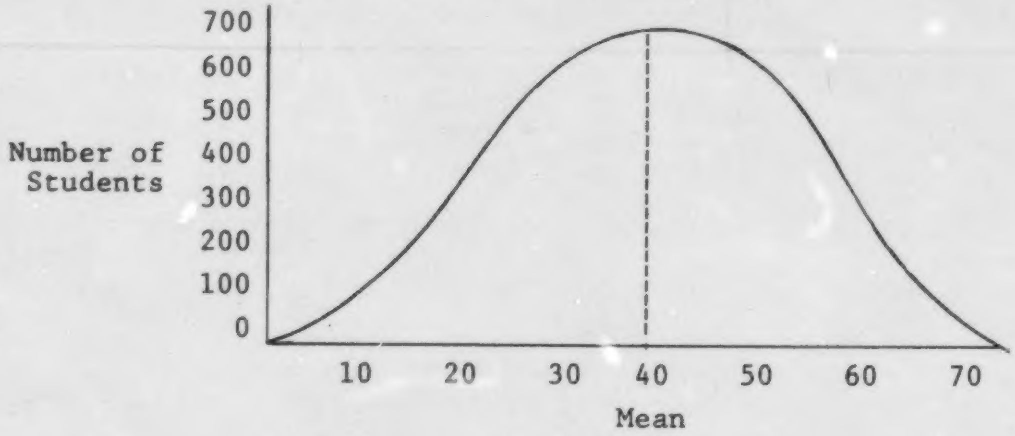
Basic to the discussion of derived scores for norm-referenced tests is the concept of the normal distribution. A normal distribution is a distribution which is perfectly symmetrical about its mean and has a bell shape. Scores are concentrated around the mean with fewer scores at either extreme. The general form of the normal distribution is shown below:



The distribution of test scores in the norm group is assumed to conform to the normal distribution and is assumed to be typical of the population to be represented by the norm group.

The distribution of test scores within the norm group is simply a summary of how students scored, that is, how many students achieved each possible score. An example is shown below. A distribution is computed for each grade level or age level, and becomes the basis for calculating all subsequent derived scores. Typically, the distribution shows that most students' scores at or near the mean, or the average score for the group, and few scores are extreme, either very high or very low.

Sample Distribution of Standard Scores



Total Test Score

Because of the performance of the norm group is basic to the interpretation of test scores, one should examine critically this group's validity. Several questions you might ask include:

- Do the students particpating in the norming truly represent the intended population? Were students similar to those being test included?
- Was there a sufficient number of students included in the norming to warrant generalization?
- Is the performance information relatively current?
- Was the norming conducted at the same time of year that the students took the test?

The answers to these questions can be found in the examiners manual for the test. Test publishers report national norms. If it is found that the answers to any of the above questions are negative, one should be skeptical about using the norm to interpret the test results.

One additional warning in interpreting test scores needs to be made. Because the scores derived from different tests are based on the performance of different norm groups, the derived scores are not directly comparable. For example, a score at the 50th percentile on the California Test of Basic Skills is not exactly the same as scoring at the 50th percentile on the California Achievement Test. Perhaps the norm group for the former test was composed of higher achievers than the latter, or vice versa. In addition, the two tests may measure different skills.

Percentile Scores

One way to determine how a student's performance compares with the norm group is to derive a percentile score. A student's percentile score indicates the percentage of the norm group whose raw scores fell below the student's raw score. For example, performing at the 60th percentile means that a student's raw score was higher than 60 percent of the students in the norm group.

If a student earns a raw score of 82 items correct out of 100 total items on a science test, this would be equivalent to the 98th percentile if 98 percent of the students who took the test received scores below an 82.

A percentile score shows how a student ranks with respect to the performance of the norm group. A percentile score ranges from one to 99 and is derived from the score frequency. That is, the number of students in the norm group achieving each possible score is computed. Percentiles are not difficult to compute. The percentile score corresponding to each raw score is calculated as the sum of the percentage of students scoring below the raw score plus one half the percentage of students scoring the same raw score. An example is shown in Table 1 (see following page).

Percentiles have disadvantages. The distortion of percentile scores around the mean can be a serious problem. All test scores are estimates, and can easily vary a point or two. Test publishers, in fact, often report an index called the standard error. This index takes into account chance errors and offers one basis for determining the range within which a student's true score probably falls. Using this statistic, a student's true score can be interpreted as within the interval bounded by the score the student received on the test plus and minus one standard error.

If the distribution is normal, a percentile difference will not represent the same amount as an equivalent raw score different. For example, the raw score difference between the 50th and 59th percentiles is not as great as between the 90th and 99th percentiles.

Table 1

Sample Raw Score and Percentile Score Equivalents

Total Raw Score	Number of Students Achieving Score	Cumulative Frequency	Cumulative[1] Percentage	Percentile[2] Score
1	1	1	1%	1
2	2	3	3%	2
3	4	7	7%	5
4	10	17	17%	12
5	24	41	41%	29
6	22	63	63%	52
7	16	79	79%	71
8	10	89	89%	84
9	4	93	93%	91
10	2	95	95%	94
11	2	97	97%	96
12	2	99	99%	98
13	0	99	99%	98
14	0	99	99%	98
15	1	100	100%	99
16	0	100	100%	99

[1] Cumulative Percentage = $\frac{\text{Cumulative Frequency}}{\text{Total Number of Students}} \times 100$

[2] Percentile Score = Cumulative percentage Below Score + ½ percentage achieving score

Percentile scores can be misleading because they do not tell you how much a student knows, but rather how the student's performance compares with others. For example, a student could score in the 84th percentile by answering correctly only 50 percent of the items. Although one might say that the student is doing well relative to other students, you could not say that he or she had mastered the test content.

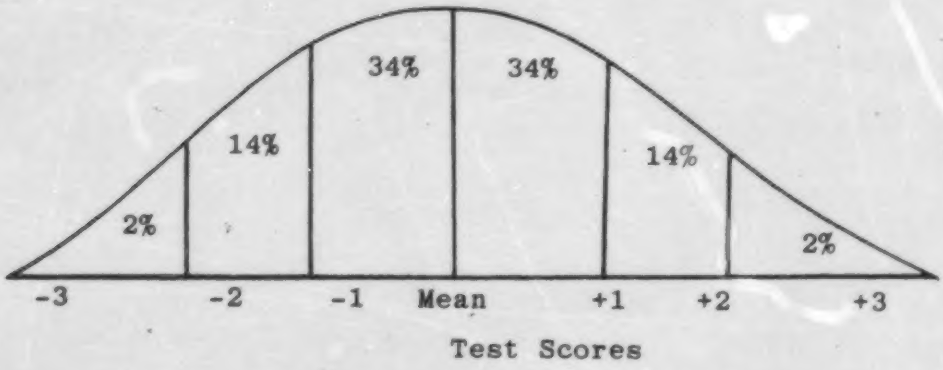
Standard Scores

Another way of determining a student's relative standing with respect to the norm group is to derive the student's standard scores. Although there are several types of standard scores, among them the z-score and the t-score, each indicates how far a student's raw score deviates above or below the norm group mean. The distance is expressed in terms of standard deviations. The standard deviations is a measure of how spread out the scores within a group are, and basically is an average of how much the scores in the group deviate from the mean.

What does it mean that a student's raw score is some number of standard deviations above or below the mean? The interpretation becomes more readily understandable by reference to a normal distribution of scores. One of the most useful properties of a normal distribution is that when it is divided into equal intervals, a fixed percentage of raw scores fall within each interval. For example, when a normal distribution is divided into intervals one standard deviation wide, 34 percent of the raw scores fall within the interval from zero to one standard deviations from the mean, 14 percent of the raw scores fall within the interval from one to two standard deviations from the mean, and two percent of the raw scores fall within the interval from two to three deviations from the mean. These percentages apply to standard deviation intervals both above and below the mean (see Figure 2).

Figure 2

Normal Distribution



To the extent that the distribution of raw score frequencies approaches a normal distribution, the number of standard deviations a student's raw score deviates above or below the mean can be given percentile score meaning. In a normal distribution, the average score corresponds to the 50th percentile. A student scoring one standard deviation above the mean would have a percentile score of 84, while a student scoring two standard deviations above the mean would have a percentile score of 98. On the other side of the mean, a student's whose raw score was one standard deviation below the mean would have a percentile score of 16.

The distribution of scores within the norm group is not always normal. The number of standard deviations above or below the mean, therefore, does not always directly translate into percentile scores. The correct standard score and percentile score equivalents should, however, be provided in your test manual. In any case, as the above discussion suggests, the number of standard deviations above or below the mean can furnish a yardstick to determine how unusual a student's raw score is. For example, if a student scored within one standard deviation of the mean, the score would not be very unusual, while if a student scored more than two standard deviations from the mean, then the performance would be quite extreme--either extremely good, if above, or extremely poor if below the mean. With this overview to the interpretation of deviation, we turn to the definition of two commonly used standard scores: z-scores and t-scores.

z-scores. z-scores tell how many standard deviations a student's raw score is from the group mean; they usually range in value from -3 to +3. A student's z-score is computed by subtracting the mean raw score for the norm from the student's raw score and then dividing the difference by the standard deviation from the norm group. For example, suppose a student scores 58 correct on a test. The norm mean is 56 correct and the standard deviation for the norm group is four. The student's z-score would be:

$$\frac{(58-56)}{4}, \text{ or } +0.5.$$

This z-score indicates that a student's raw score is 0.5 standard deviations above the mean of the norm group. A student who has a raw score of 46 on the same test would have a z-score of:

$$\frac{(46-56)}{4}, \text{ or } -2.5.$$

This score indicates that the student's raw score is 2.5 standard deviations below the norm group mean.

t-scores. A student's t-score is derived from the student's z score. It is the result of multiplying the student's z-score by 10, and adding 50. For example, a student who has a z-score of .05 on a test has a t-score of 10 (0.5) + 50, or 55; while a student with a z-score of -2.5 on the test has a t-score of 10 (-2.5) + 50, or 25. Here, then, in contrast to the z-score scale where the mean is one and the one standard deviation is equal to one, the mean of the t-score is 50, and one standard deviation unit is 10.

t-scores indirectly indicate the number of standard deviations a raw score deviates above or below the mean. A t-score of 70 indicates that the student's raw score is 2.0 standard deviations above the mean, a t-score of 35 indicates that the student's raw score is 1.5 standard deviations below the mean of the norm group raw scores. Given a t-score, one may determine the number of standard deviations the student's raw score deviates above or below the mean by deriving a standard score which indicates this directly, that is, by deriving the student's z-score. Since a t-score is equal to 10 times a z-score plus 50, a z-score is equal to a t-score minus 50, divided by 10. So, given a student's t-score of 70, the student's z-score is $\frac{(70-50)}{10}$ or +2.0. And given a student's t-score of 35, the student's z-score is $\frac{(35-50)}{10}$ or -1.5.

The advantage of t-scores over z-scores is that they avoid negative numbers and decimals, which makes calculations easier. It should be noted that many other standard scores exist. For example, the Scholsatic Aptitude Test uses a standard score scale where the mean is 500 and one standard deviation equals 100. Normal values for the scale thus range from 200 to 800.

Stanines

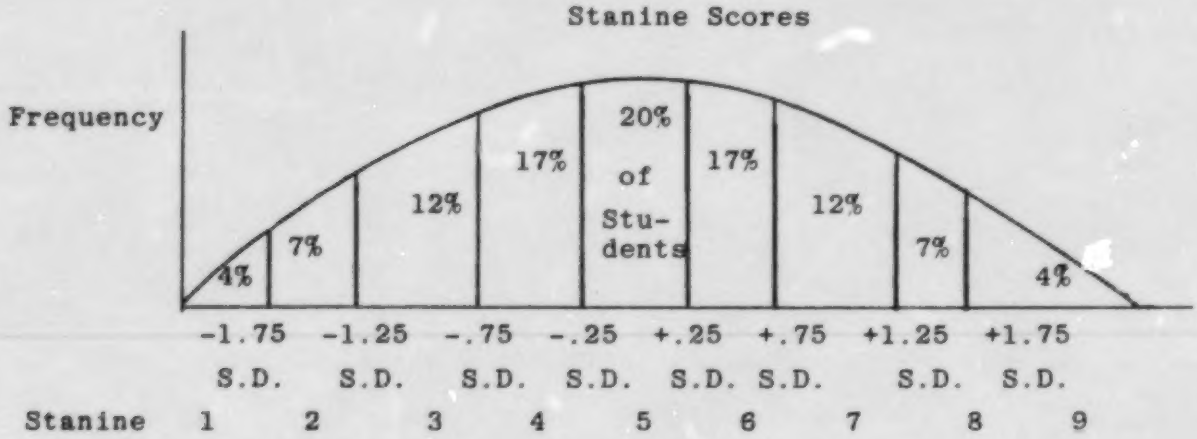
Stanines are derived scores with a mean of 5 and range from 1 to 9. They divide a normal distribution in nine parts.

Stanine scores, like standard scores, indicate how far a student's raw score deviates from the norm group mean. The distribution of raw scores in the norm group is divided into nine intervals. The inner seven intervals (stanines 2-8) are one-half standard deviations wide, and the outer two intervals (stanines 1 and 9) are greater than one standard deviation (see Figure 3). Stanine 5 straddles the mean and contains all raw scores within 0.25 standard deviations on either side of the mean.

The remaining stanines are evenly distributed above and below stanine 5. Stanines 6 and 4 contain, respectively, all raw scores 0.25 to 0.75 standard deviations above and below the mean, while stanine 1 contains all raw scores more than 1.75 standard deviations below the mean. When a normal distribution is divided into stanines, each stanine contains a fixed percentage of raw scores: 20 percent of the raw scores fall within stanine 5; 17 percent of the raw scores fall within stanine 6 and 17 percent of them fall within stanine 4; 12 percent of the raw scores fall within each of the stanines 7 and 3; 7 percent of the raw scores fall within stanine 8 and 7 percent of them fall within stanine 2; and, finally, 4 percent of the raw scores fall within each of stanines 1 and 9.

Figure 3

Normal Distribution



Many have recommended the use of stanine scores rather than percentile scores. Because stanines cover a range of percentile scores, they tend to be more stable estimates.

Normal Curve Equivalents

Normal curve equivalents are a relatively new derived score and have been used in ESEA Title I evaluations. Like percentiles, normal curve equivalents range from 1 to 99, with a mean of 50. Normal curve equivalents, however, have a standard deviation of 21.06 so that normal curve equivalents of 1 and 99 correspond to the 1st percentile and 99th percentile, respectively. One disadvantage of normal curve equivalents is that they can be easily confused with percentiles.

Grade and Age Level Equivalent Scores

Grade- and age-level equivalent scores attempt to tell where a student's raw score falls with respect to the average performance of students at various grade or age levels. The average raw score, the median of students at each grade or age level, then defines the grade (or age) level equivalent for the test. Grade or age level equivalents for grades that were not tested are computed by interpolating based on the trends in the data.

If a student's score on a test was the same as the median score for all beginning second graders, then the student's grade equivalent score would be 2.0. If the student scored the same as the median score for all beginning third graders, a grade equivalent score of 3.0 would be assigned. As mentioned above, interpolation would be used to assign grade equivalent scores in between.

Although grade- and age-level equivalent scores have great intuitive appeal, they suffer from a number of methodological problems. A primary problem is the way scores are interpolated, that is, how scores are derived for levels not tested, and how scores between tested levels are computed. Using the example above, what does a grade equivalent score of 2.7 mean? Or, what does a grade equivalent score of 4.6 mean if the test was not given to students above the third grade? Interpolation and extrapolation are imprecise.

Also, small sampling errors can be compounded into large errors in extrapolation and then make grade equivalent scores very misleading and inaccurate.

In addition to methodological problems, age and grade equivalent scores are often misused. If a fourth grade student obtains a grade-level equivalent score of 7.0 in mathematics, it does not mean that the student can do what a seventh grader does. It only means that the student got the same raw score on the test as the average seventh grader participating in the norming or that the score was obtained through extrapolation. A grade-level equivalent score says nothing about the content a student knows. The items the fourth grader answers correctly to obtain a seventh-grade equivalent score may be quite different from the items the seventh grader answers to obtain the same score. The test given the fourth grader most likely does not include many of the skills or content that would be expected of a seventh-grader.

Another common misuse of grade- or age-equivalent scores is to use them as standards and assume that all students should be performing at least at their own grade level or age level. Given the way these scores are calculated, one can expect half the students at any age or grade to fall below the chronological equivalent.

Because of the methodological problems of grade and age equivalent scores and their frequent misinterpretation, many prominent measurement experts have suggested that these types of scores should not be used.

Chapter V

APPLICATION OF STANDARDIZED TESTS TO YOUR PURPOSES

In the previous chapters, information on the various kinds of tests was presented. Factors which should be considered in the selection of tests were discussed, and in the last chapter the various results of tests were presented. In this chapter, the application of the test scores to classroom and district use is presented. That interpretation is based on the question of "How are the results of the tests to be applied to the improvement of instruction?" In seeking to answer this question other questions may develop. What does a district release to legislators, citizens, parents about the performance of students on tests? These and other questions about what information is necessary to release or use may not be provided expressly by the test results. To the uninitiated it might seem shocking that fully half of the students in the district scored below the fiftieth percentile on a test. If the parents of the district would be concerned about such information, then someone in the district must make an effort to educate the parents or others drawing similar conclusions. This is certainly one of the responsibilities that accompanies testing. Persons associated with and concerned with the testing of students will want to know the results. Before addressing their concerns first see to it that the teachers have the information necessary to digest the results and apply them to instruction. Teachers are the front line contact with students and parents, and as such have priority for test information.

An additional responsibility to correctly assess the resources and time necessary to evaluate test results and incorporate them into the education process exists. Some person in the district must be charged with the responsibility for accomplishing this task. Those involved need to know what is available, and what is expected of each participant in the testing program. Decisions about testing should be carefully weighed in light of the information developed in such an analysis. The reactions of various interested groups should be considered so that a concensus is reached or lacking a concensus, a policy is adopted with authority to implement.

Careful monitoring of the testing program should be provided for in the implementation of the policy. Users and those seeking information regarding performance can abuse the results of the best, most carefully adopted test.

There is a demand for testing, but a considerable effort is required to properly implement the program. It is a responsibility not to be taken lightly.

Educating the Test User

With this charge for responsibility in mind interpretation of test scores can begin. How much effort will be expended to get the test results to the people who will be using them? Someone with a realistic idea of how much time will be required for teachers and administrators to process the results of the test should make such estimates in advance of the test selection. Then factors affecting the processing should be explored. In the best teaching form, it should be gone over and over again until it is right. While the school year is in progress is a good time to assess the level of understanding held by the staff. Those who would benefit by courses or inservice work on test use and application should be given the opportunity to improve their skills. Information about the test selected and its advantages and properties should be available at the school, the union office, teacher centers--wherever three or four are gathered--so that the widest access possible is available.

There is absolutely nothing wrong with offering incentives to teachers to learn about the test they will be using. Money is nice. So is released time, credit toward continuing education hours, etc. If as much money were put into getting people to properly use tests as is spent on the tests themselves, it would still be less than one percent of the per pupil cost. In these times of inflation this is an exceptional bargain. Remember that the decision about how enthusiastically people will be willing to work on the testing program or any other part of the school program needs to be a joint decision involving teachers, administrators, and board members. It is about time we tried a way that responsible people could agree might have a chance for success.

Once the amount of effort and resources that will be put into the testing program is decided, some other things have to be reviewed. What was it that the district decided was their purpose for testing in the first place? In conjunction with this some restatement of the limits of the test selected should be made. The format for presentation of the data, either hand scored or computer generated, should be identified. When and where will the data be available and to whom? The answers to these questions can be presented in fairly simple and direct terms. Plain talk is the key to getting people to listen. Some group in the district, perhaps the persons who served on the selection committee, should prepare all of the information that is necessary for presentation of the test to the staff of the district. Someone in authority should see to it that provision is made for the staff to receive the information. The following are possible considerations that might be made in the implementation of a testing program.

The chief officer of the district responsible for testing should have some procedure available to get the information to teachers. This would not preclude an agreement with the union to have meetings to offer inservice during the school day or at some other time agreeable to the staff. Participation of teacher representatives on the test selection committee as well as a survey of needs during the selection process will produce the kind of enthusiastic cooperation necessary to make the program a success. The preparation of the information which will be presented should include statements as to the district's purpose for testing. If the district sought to have achievement tests so that generalizations about developed abilities could be made, then it should be presented in that way. The basic information available to teachers should include at least a manual for the test and comments by the committee about the appropriateness for use in the district should be made. The district may want to indicate that the test was checked for reliability and validity. Statements about how well the content of the test matched the curriculum and supplementary materials used on level in a majority of the classes in the district would be appropriate. Specifics could be included in reference tables without bogging down the reader. It would also serve as a check for teachers as to what was being taught in the various classrooms of the district. A comparison of the basic skills and content covered by the test could be contrasted with the curriculum used in the district. What was tested that was normally taught and vice versa would be information that could help relieve some of the expectation of teachers for test scores. It could also serve to point out the areas that were usually taught that were felt by the district to be important enough that they needed emphasis. Such a summary comparison does not relieve teachers from the responsibility of examining test items to determine if there is content validity for their students. If lack of content validity is noted some procedure for notifying the district should be specified.

Some discussion of the norming group and the characteristics of the district could be included. Comments about the significant differences would serve to point out what reasonable expectations should be made about students' performance on the test. Special groups of students or student populations might be identified so that teachers of those students will not be caught unawares by the group's performance. For example, if students with limited English abilities are predominant in some schools or classrooms, then knowing the advantages and limitations of the test for a particular group will allow teachers to adjust their expectations as necessary. This will help with the morale problem that accompanies unexpectedly low results.

Teachers should know the format for the score reporting. They should also know if others in the district will be reviewing individual or group data about the students so that duplication of effort can be avoided and proper attention to the scores can be given. Teachers should not find out from the headlines in their local paper that the class they are teaching is far below the district or national average. Such information should be first presented to and evaluated by the teacher as suggested previously so that suggestions for improvement can be included in any reference to district scores. At no time is it appropriate to use the scores of one school to compare the scores of another school. Such comparisons must be made in relationship to some standard that has meaning. The differences for each school over time must be considered. The scores are best presented in relationship to the facts of the situation.

The same goes for comparisons of student scores with the particular teacher or class. By their nature, the standardized achievement tests must test knowledge learned in previous grades or years as well as the current one. The notion that arbitrary goals of improvement can be set without considering where the student began is ridiculous and wrong. Test scores are not intended for such purposes. Appropriate use of the scores might include a review with the teacher of the strengths and weaknesses of the individual student as shown by the test scores and other measures such as grades, or by personal observation. Allowing for error and related factors, the generalizations that can be made should serve to improve and focus the instructions where possible. In reviewing the test scores, students may score high on general reading, vocabulary, listening skills and auditory portions of the test. If for the same students markedly different performance (two or more stanine scores difference) on the math or science is observed, it might be attributed to some problem other than comprehension of the vocabulary on the test or lack of reading skills necessary to complete those sections correctly. Teachers should be encouraged to look to the test items and determine what kinds of skills and content were missed in class. Math computation might have been presented in a vertical format but on the test only horizontal format problems were included. This unfamiliarity with the format might be disguising developed abilities. This kind of item review would produce some estimate of the source of the problems. By observing the problems the teacher should be able to indicate the resources necessary to overcome the deficiencies. The support for the teacher's recommendations should be followed and acted upon. If such recommendations are not met, the district must reconcile itself to the fact that limited change is possible.

It is important to remember that the tests are inexact and that some things beside student performance or test characteristics may affect scores. If abnormally low scores are discovered in a particular class of students, some check on the administration of the test would be in order: a review of the conditions of the class, the instructions given, the physical conditions of the testing room, and similar non-test or student factors should be considered. Unusual attendance, activity, illness, or acts of nature could account for the differences. Even though tests are supposed to be able to accommodate some variations from the normal situation they may not tolerate extreme conditions.

Once teachers have the information in hand about the test, it would be useful to have someone who is not a direct supervisor or evaluator available to answer questions. The press of normal school business is such that counselors, principals, and consultants are not always available to spend the necessary time to fully discuss the questions raised. Some teachers may fear that they will appear less than competent if they raise a particular question. If such questions remain unanswered, serious errors in interpretation of test scores may occur. Designate someone who might be available after school hours or during the planning periods of the school day for some period of time prior to returning the test scores. Teacher centers or the union office could also provide information to those with questions. These kinds of additional resources should not be overlooked in the presentation of test information or the clarification of questions.

The directions about the use of test scores should include some information about how the scores will be arrayed. The available methods of computer printing and analysis can eliminate much of the fatiguing busy work previously performed by the classroom teacher. Where available, this option should be chosen. Samples of the scores which will be presented along with the forms for analyzing the data should be part of the presentation of information. The class record, a summary of all student scores and norm information, is a valuable tool for teachers to use. Item review to identify those skills or content areas most often missed should be included as well. The manual for some tests will provide sample worksheets and forms which can be altered to fit the district's needs. Uniform preparation of these kinds of analysis sheets is imperative and teachers should be told clearly what is needed in the way of cooperation on this matter. While the test manual may be helpful, the addition of the local interest in a successful testing program cannot be ignored. Teachers must have a manual and a test to refer to so that they can begin to examine the items for form and content. Teachers need to be familiar with all aspects of the test if they are to gain confidence in the test and begin to use it to supplement other criteria in forming educational decisions.

When scores are returned to the district, the distribution of the scores and the accessibility to them by teachers is important to allow for. Careful planning as to the location of the scores and any security necessary to protect students from unauthorized use of their scores must be thought through. Time to assimilate and prepare the scores in a meaningful way must be designated and provided. When the scores are distributed to the school are they tucked away in the counselor's office or can teachers get them from another source? Teachers should know when scores are available and how to get access to them. Too often in the past the test scores have been returned too late to be of value to the teacher who has tested the pupil. This is essentially true in spring testing situations. In fall testing, the scores often come back too late to use or incorporate into planning for individual student programs. The district should clearly state precisely what it hopes to have the teacher use the scores for and what timetable is to be followed.

Available Test Scores and Their Properties

In the previous chapter, the kinds of scores reported were introduced. Scores are available in several forms based on the raw score performance. Scores may be expressed as raw scores (the number of items answered correctly), percentile ranks, stanines, grade equivalents, and scaled scores. Some test publishers may use still other means to report the scores. In any case the raw scores are the one thing that tests have in common. How these raw scores can be transformed into other comparable units of measure is something that should be specified by the district and should match the intent of the publisher.

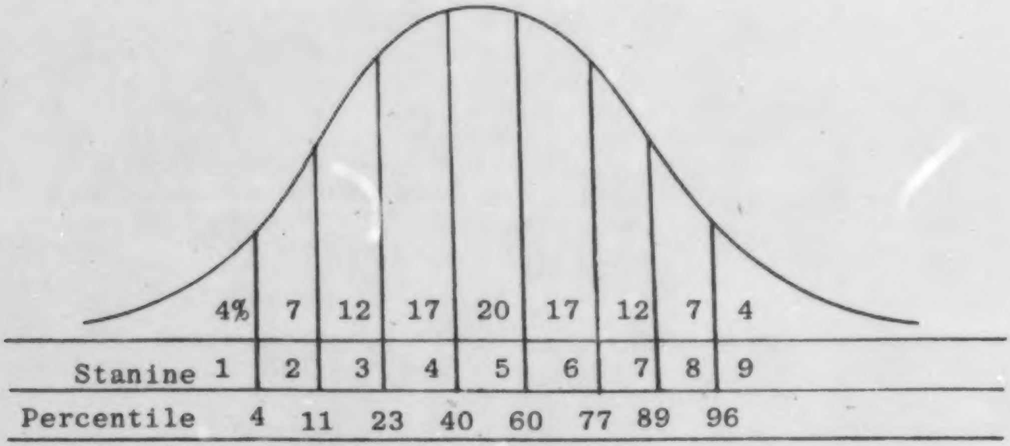
Percentile Ranking

If percentile ranks are to be used, teachers should understand that the percentage of cases in a distribution at or below any given scores value determine the percentile rank of that score. Since percentile scores do not provide equal units of raw score measure it would be helpful to remind teachers that near the center of the distribution the scores will bunch up. The difference of a few raw score points may make a large difference in percentile ranks at the middle of the scale and a small difference near the ends of the scale. An example is provided in Figure 5.3 and what the effect of this feature of percentile scores does is illustrated. Averaging percentile ranks is difficult because of the difference in the measure of value for each unit and interpretation of the magnitude of difference between percentile scores is difficult. This is not to say percentile ranks are useless, but this characteristic should be noted and accounted for by the teacher. The publisher should identify percentile ranks for the various times of the year in which tests may be given. A raw score for a fall administration will yield a different rank for the spring administration, so the correct table should be identified and used by the teachers.

Stanines

Scores may be reported in stanines also. Stanines are groups of values on a nine point scale of normalized values. A ranking of one is the lowest, nine the highest, and five represents the average performance for pupils in the norm group. They are tied to the percentile ranks for a normal distribution and can be obtained directly from the computation of percentile ranks. Since the stanines are equal units the bunching effect of percentile ranks is avoided. Differences in stanine ranks are comparable, with a difference of six and eight being similar to a difference between four and six. By using stanines, teachers can avoid making distinctions which are too fine to be accommodated by the test. The relationship between stanines and percentile ranks is illustrated in the figure below:

Figure 5.1



By observing the differences in stanine scores on subtests of batteries, the teacher can identify those areas in which a student is excelling or having difficulty. For most cases a difference of two or more stanines in a score between tests is said to be cause to examine the performance in the exceptional area. One problem with the use of such measures as percentile and stanines is that change over time is not usually identifiable.

Grade Equivalent Scores

Grade equivalent scores will not be much help in showing advancement over time either, unless longitudinal data is available. Even then different students will perform differently, with only those near the median showing an average growth of one year for each school year of education. Grade equivalent scores are determined by first translating all spring and fall raw scores medians for a test at each level into a common level. These raw score points are plotted on a graph and a line is drawn that best fits the points. It is possible that only a few points will be identified and other points are assumed to fall on the line of best fit. These points along the line are called grade norms. The raw score corresponding to any grade equivalent indicates the appropriate score that would be made by the pupils in the standardization programs at a specific point in the grade. If a test is given in a district that begins earlier or later than the frame of reference for the test standardization, then teachers should be notified and the necessary adjustments made.

Grade equivalents at best describe how a student at a particular grade level would do if he took the test for the level tested. For example, if a test is given to fourth graders in October (designated by 4.1) and a seventh grader performing at the average level for seventh graders took the test, the score derived would be 7.1, indicating October in the seventh grade year. If a fourth grader took the test and received a grade of 7.1, it would mean he achieved an identical score, not that he would be able to perform math like seventh graders. Computation and division and other content areas normally covered in the higher grades would not have yet been taught to the fourth grader. It would be inappropriate to place him out of grade level. In fact, any score that is a grade equivalent more than two years beyond the level should be referred to as well above average and to attach more significance than that would be misleading.

Scaled Scores

The score that may be the most useful in comparing students' progress over time is the scaled score. For tests on all battery levels, the scores may be comparable as in the case of the Metropolitan or the Stanford Achievement Tests. The publisher should identify this property of the derived scores if available. Once raw scores are converted to scaled scores, the battery level and the form can be ignored in further interpretations. Batteries of the test are equated and forms are made equivalent in going from raw score to scaled score. Features of the scaled score allow comparison of achievement over time and for this reason might be more useful to a teacher who is trying to generalize about the student's developed abilities than other kinds of scores.

In order to compare one test to another, for example a math test to a reading test, the scale chart provided by the publisher must be consulted. There is no direct comparison without such consultation. In addition, the scores have no meaning by themselves and cannot be used in interpretive decisions. Percentile ranks and grade equivalent scores offer more in this activity.

Raw scores

- o Convert raw scores to derived scores such as percentile ranks, stanines for later use in comparisons.
- o Use in statistical analyses when computing correlation of coefficient and similar procedures.

Percentile ranks

- o Use to compare pupil's standing on a test or ranking in relationship to a national or other group standard.
- o Use to compare results among test batteries.
- o May be a choice for reporting test results to parents, pupils, and others who are not familiar with testing and measurement.

Stanines

- o Same as percentile ranks plus they may be used for making comparisons with some other variable in performance such as general learning ability.

Grade Equivalents

- o Use for interpreting performance of groups such as an entire class or grade.
- o Use for measuring advancement over time when longitudinal data is available and relative level of achievement is accounted for in data.
- o Use for determining relative individual achievement when consideration is given to the differences that may be associated with high, average, and low achieving characteristics of the student.

Scaled Scores

- o Use to study achievement over time as data is collected and reviewed.
- o Use for interpreting results when testing is out of level.
- o Use for most applications when conducting statistical analyses.
- o Use to compare different forms and batteries of tests.

Using Test Scores

Finally all of the education and preparation will be completed and the scores made available to the classroom teachers. Before beginning to record scores onto a summary sheet it is useful to see if the scores seem right. If based on the knowledge of the teacher about the student's past performance the scores are generally what was expected, then further interpretation can be pursued. Stanine scores for students are useful for such initial examination because of the ease with which assumptions about "above average, average, and below average" categories can be made. If a student is scoring in the second stanine on math portions when you realize the student to be generally superior in math, then the test performance should be questioned. The fault may lie with the content, presentation format, or other variables. The physical condition of the students, the administration of the test, confusion on the marking of answer sheets might account for the marked difference observed. If many students score lower than anticipated, perhaps it is due to a lack of coordination with what was taught and what was tested. This should be checked against the validity information developed by the district test selection committee. The match between curriculum and content should have been identified and adjustments suggested. Test publishers will often identify the content of the test items in the technical manuals. These sources should be consulted to determine the extent of a match in circumstances that cannot be attributed to individual student performance.

Sorting

The apparent sorting of students into categories and smoothing of individual differences is a function of test interpretations of this type. Critics of testing will decry such sorting and labeling as damaging to students. A few statements in support of sorting are appropriate and offered for your consideration. Sorting is neither good nor bad without reference.

- While teachers may be aware of the differences between twenty eight or thirty students, in practice these differences are minimized and groups are formed either consciously or unconsciously as teaching occurs.
- The time constraints on teaching a course, limitations of books and materials available, and the physical characteristics of the classroom all lend themselves to sorting. To some extent efficiency of effort requires such sorting.
- Sorting is appropriate when making some decisions about presentation of information and can lead to efficient teaching methods which allow teachers to be more flexible and individualize instruction with the remaining time and resources.

- Not everyone's score on a test is an indication of performance ability. Terms developed in the sorting should be applied to the context in which they were developed. Knowing you were below or above average is meaningless in most applications unless a relationship to a standard of either a group or the individual is expressed concurrently.

- The performance level in relationship to a group standard or individual standard can help to identify an appropriate activity for the student in which the greatest opportunity for educational achievement can occur.

Preliminary Activities

If a review of the scores shows that most students are at the high end of the scale it is likely that another form of the test should have been used. If everyone scores in the upper ranges, the differences are masked and the test was probably too easy to be of use in determining strengths and weaknesses. Out of level testing might be pursued to gain more definitive information.

References to the level of ability of the school or class from previous testing or other measures for student ability may help to identify the student who is achieving out of sync with previously demonstrated abilities. Not only should the lowest achieving or the highest achieving students be observed in light of school performance or previous history, but all students should get consideration. The students in the middle stanine ranges may well be some of the ones that could have been expected to achieve in the upper ranges. Too often these bright students are missed when their performance is substandard for their ability. Students who score well above what you expected may be exhibiting a perfectly natural behavior in the structured testing situation and have an entirely different set of behaviors in the regular classroom setting. The initial review can help establish a level of confidence in the test results, so that further specific interpretations of the scores can be made.

A relationship between reading ability and test performance can be expected. Reading ability is required for many of the subtests and a student who scores low on reading and high on math computation may be delivering a message about needing help in reading. High reading and math application scores (thought problems) followed by low scores on math computation subtests may indicate a need for math improvement in the computation area. If math questions that rely on reading ability can be handled successfully then the problem lies more likely in the student's math computation skills.

Student Scored Summary

The individual pupil record form provides teachers with the information from which most other observations will be made. Figure 5.2 is a sample form that might be presented in reporting student scores.

Figure 5.2 Student Score Summary

Name Larry Rudner Grade 3

Teacher Anne Goldblatt Date of testing 10/17/80

School Ward Elementary District Unionville State Mass.

Test	Number Possible	Number Right	Scaled Score	Grade Equiv.	Percentile Rank	Stanine		
Reading	55	39	633	2.9	44		4 [✓ 5] 6	
Math	45	26	493	2.7	32		[✓ 4] 5 6	
Language	55	42	537	3.0	66		4 5 [✓ 6]	
Science	40	21	421	1.9	18		4 5 6	
Social Studies	40	27	516	3.0	50		4 [✓ 5] 6	
Basic Battery (R & M & L)	155	107	549	2.9	44		4 [✓ 5] 6	
Complete Battery (Basic & S & SS)	235	153	506	2.7	40		4 [✓ 5] 6	

The raw scores showing the number correct on each test are entered in the appropriate space, here labeled "number right." If hand scoring is done by the teacher, a key is usually provided which allows a quick visual check to determine if the correct answer is marked. These are totalled for each test and generally identified as the raw score. While marking the booklet or score sheet it would be useful to identify which choice was made by the student in each wrong answer. This information used later can assist in determining something about the choice in relationship to behavior or abilities. For example, a student who consistently marks the answer on a math test which has the decimal misplaced may need review or emphasis on decimals. To know only what has been seen as correct offers a limited view of the student's performance.

Once the raw score is entered and the totals of the basic and the complete batteries tabulated, the raw score can be converted to a derived score. The derived score may be calculated as demonstrated in earlier chapters of this manual, but the most practical way to proceed is to use the tables generally found in the instructor's manual accompanying the test.

When using the tables in the manual, be certain that you have the correct form of the test and the correct norm for the time of year in which the test was administered. Also be conscious of the fact that many of the derived scores are based on mean or median performance and are not in and of themselves indicative of desired achievement levels.

In the example given in Figure 5.2, three derived scores will provide some insight into the performance of the student as shown on this test. From the percentile rank column, notice that the student was in the 44th percentile on Reading and the 50th in Social Studies. These two scores are fairly close to the center of the distribution of scores and will be considered later to determine if they represent significant variations in performance or are a function of the "bunching" of scores that sometimes occurs when using percentile rank. The Science score at 18 is in the lower third of the scores and the 66 in Language is at the upper third dividing point. The assumption is that the student is better in language skills than in math skills, but more importantly you can see the highest and lowest score ranges. Notice also that the difference between the highest score and the others is similar to the difference between the lowest score and the others near the center. Both Math and Language scores should be examined more closely to see what relationships might exist. By looking to group performance information and other student records, some estimate of how close to expectations the student performed can be determined. In the case of the low science score it might be to the teacher's advantage to note the score and later analyze the specific items missed. If specific concepts were taught in the curriculum, the requirements for reading and vocabulary on the test were different, then in order to do reasonably well on the test some changes in the curriculum content might be made. An alternative would be to find another test for science or a battery of tests that has content closer to the curriculum. The administration of a locally developed objective or criterion referenced test might be a useful beginning point.

Percentile Midpoint Bunching

A review of the scores near the midpoint of percentile rankings can be of value to determine if a wide number of scores are separated by only an answer or two near this performance level. From the tables provided by the test publisher, the following can be observed. The Social Studies score of 27 produced a percentile ranking of 50 for this particular test. The other scores produced percentiles as shown.

Figure 5.3

Relationship of Percentile Rank to Raw Scores

Social Studies Raw Score N Correct of 40	Scaled Score	Percentile Rank	Difference from Student's actual Score
24	474	34	Score -3
25	488	40	Score -2
26	502	44	Score -1
27	516	50	Student Score
28	530	54	Score +1
29	544	64	Score +2
30	558	70	Score +3

By marking one answer differently, the percentile ranking would have changed upward to the 54th percentile or downward to the 44th percentile. A change of two answers would produce a rank change from 10 points lower up to 14 points higher. Three answers different would have moved the student from the middle third to the upper or lower third of the rankings.

The effect of a one to three point raw score difference is less drastic for the total battery. In this test the percentile ranking of 40 for the whole battery is affected only two points upward for three additional correct answers and four points downward for three additional incorrect answers. At the midpoint in the scale three answers either way produce a change of plus or minus four points. Such properties of percentile rankings should be kept in mind as the scores are reviewed. Decisions made based upon percentile rankings should account for the swings produced by a few answers at various points in the scale. Generalizations rather than tightly drawn conclusions are better uses of the scores in this instance.

Cluster Analysis

The use of cluster analysis may help determine the specific strength and weakness of a student within a subtest. Often the publisher will provide information about the national median performance on the items in a cluster of related items on the test. Figure 5.4 represents the mathematics cluster on which a student scored 26 correct answers out of 45 possible. For this exercies the number in the lower right hand corner of each box will represent a norm referencing to the national median performance. A discussion of criterion referencing will follow. The number in the upper left hand corner is for the student's score on the items in that cluster.

Figure 5.4
Cluster Analysis for the Math Sub-test

Mathematics

Item	9 / 7 Numeration	7 / 6 Geometry & Measurement	4 / 7 Problem Solving	10 / 10 Operations: Whole Numbers
No. Possible	10	10	10	15

Referring to the manual a table appropriate for the grade level and time of year for test administration showing how the student did in relationship to the average for the norm group can be made. Also some indication of how difficult a cluster of items was for most students can be seen. If the average number correct was four of ten it would likely be more difficult an item cluster than one in which students answered eight of ten correct. In the example in Figure 5.4, the student's strengths and weaknesses shown in the cluster indicate that the student is not evenly skilled in the various operations of mathematics. In both Numeration and Geometry and Measurement the student was above the national average. Problem Solving was particularly difficult for the student. The score was just over half of the average score for all students taking the test. Assumptions on general skills within the subtest can be made by the teacher providing that the items in the cluster match closely the items covered in the curriculum. By keeping track of the performance of various students in clusters some indication as to whether individual attention must be given the student is shown or perhaps other changes in the curriculum or teaching methods would be in order.

Criterion referencing can be applied to the same kind of cluster analysis by substituting the teacher's criteria or some other criteria for the norm referenced scores. If a teacher felt it was important for a student to pass all items in the Numeration cluster because it was an essential skill for later instruction, then the criteria might be equal to the number possible, in this case ten correct out of ten. If the teacher felt that the minimum acceptable performance level was 80 percent correct, then a score of 8 in the problem solving cluster shown in Figure 5.4 would be the goal.

Certain kinds of tests provide results which are more appropriate for determining detailed information about competence on specific learning objectives. The Reading, Mathematics, and Language tests would be usually expected to fall into this category. By studying the responses of students to the questions some indication of strength and weakness for the pupil or class can be determined. More importantly, a study of this kind can indicate where more diagnostic assistance is needed or where some judgments will be required about how to proceed with the student. To assist in the time consuming task the option of computer assisted scoring would be useful provided such a service is available. Certainly one would not select a test that was invalid or unreliable because of the scoring feature. If tests are comparable with the exception of the scoring service, then those who would be inclined to use item analysis should look seriously at a test with this feature.

Item Analysis

Item analysis involves identifying individual student responses on a test and determining how many correct responses were given as well as what alternatives were chosen. A chart could be constructed with the objective of the test to be examined listed with other items identifying information. Additional information about the percent of students getting the answer correct could be shown for the class, the school, the district, and the nation depending upon the universe the teacher is using for comparison purposes. The responses for the choices, including the choice to omit an answer, can be summarized for class performance analysis as well as for noting individual pupils' responses.

Such a process will take some time to accumulate or to consider fully even if computer scoring is used. Hand scored tests can be summarized in the same way but it is a lot of work, especially if all students and all responses are transcribed onto a worksheet. The teacher may be interested in a sample of questions and could easily eliminate those which do not appear to be of use in examining the course objectives. A percent correct column could be used to determine how difficult an objective was for all students. Test publishers will often assign the percent correct identification to test questions to show that a specific number of students were successful in selecting the correct answer. If the percent correct is high it can be assumed that the question was of a low difficulty level and likely was designed to identify those students at the lower end of the scoring scale who need assistance in this skill area. The questions with a low percent correct number will probably help the teacher differentiate between the top scoring students and assist with identifying those who need additional out-of-level testing to more precisely identify strengths and weaknesses.

Consistent incorrect answers within clusters of objectives may flag students who will need additional help in an area. If the entire class scores below the average it may mean that this objective needs to be looked at in terms of the local program. While the test may not meet the specific criteria for the class in the way it covers content, it certainly should be some indication of the need for the teacher to look further as to the cause for the wide differences. Generally it would be reasonable to identify those clusters of items which ten percent of large groups of students have scored lower than average, a fifteen percent difference for groups of less than fifty students. The test manual would likely indicate the level of significance for these various score variations and the teacher should be guided by their specific instructions.

Group Data

The use of group data can assist in making general statements about class performance. Group data can be summarized on class record sheets in which the scores for an entire class are presented. The class record form can be developed locally. If so it should include the information useful to the district in its long term longitudinal studies. Generally test publishers may provide a suggested form or may be able to provide one with its scoring service. In any event the concern of the user should outweigh the convenience for the publisher.

covering all tests in print and all out-of-print tests once listed in *MMY*, a name index to authors of over 70,000 documents (tests, reviews, excerpts, and references) in the seven *MMYs* and *TIP II*, and a scanning index for quickly locating tests designed for a particular population.

ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service, Princeton, New Jersey, 08540.
Test information, bibliographies are available through the ERIC Clearinghouse and documents can be purchased through the ERIC Document Reproduction Service (EDRS), Computer Microfilm International Corporation, P. O. Box 190, Arlington, Virginia, 22210.

Gronlund, Norman E. *Measurement and Evaluation in Teaching*. New York: The Macmillam Company. 1965.
A good basic testing text for classroom teachers that includes simple, concise and straightforward discussions of most of the major issues covered in this manual.

Mehrens, William A. and Irvin J. Lehmann. *Standardized Tests in Education*, Third edition. New York: Holt, Rinehart and Winston. 1978.
Useful chapters on reliability, validity, and reviews of some of the more commonly used standardizes tests.

Northwest Regional Educational Laboratory. *Guidelines for Selecting Basic Skills and Life Skills Tests*. Portland, Oregon: Clearinghouse for Applied Performance Testing, Northwest Regional Education Laboratory. 1980.
Given the important role test play in education, it is crucial that test users understand the fundamental principles of proper test use. These guidelines present some of those principles, focusing specifically on the selection and purchase of published basic academic skill and life skills tests. However, though the guidelines focus specifically on tests of basic and life skills, the principles presented here can be applied to review, selection, and purchase of most achievement tests intended for use in educational settings. Aptitude tests--those intended to measure a student's capacity to learn--are not covered here.

To supplement the guidelines and further assist educators' test review and selection, the appendices contain extensive lists of currently available basic skills tests. Information is presented on test characteristics, publishers, and sources of additional, more detailed information. Although these lists are intended to be quite comprehensive, inclusiveness is not claimed. Readers are urged to consult the reference documents cited in the appendix for more comprehensive listings.

MULTISUBJECT ACHIEVEMENT BATTERIES

Tests and Subscores	Grade Level(s)	Publication Date	Publisher	Reference
CIRCUS, Levels C & D	1-3	1979	AW	NOT Sept 79
Reading				
Mathematics				
Writing Skills				
Listening				
Phonetic Analysis				
Oral Reading				
Say and Tell				
Do You Know				
Think It Through				
Things I Like				
Educational Environment Questionnaire				
Comprehensive Tests of Basic Skills Expanded Edition Forms S&T (CTBS)	Kindergarten-12	1976	CTBS	MMY 12
Reading				
Mathematics				
Language Arts				
Reference Skills				
Science				
Social Studies				
Criterion Test of Basic Skills	Kindergarten-8	1976	ATP	MMY 14
Reading				
Arithmetic				
Diagnostic Skills Battery	1-8	1976	STS	TCB Jan 77 pg. 5
Reading				
Mathematics				
Language Arts				
Iowa Tests of Basic Skills Multi-level Edition Forms 7&8	3-9	1978	HM	NOT July 79
Reading Comprehension				
Mathematics Skills				
Language Skills				
Work-Study Skills				
Vocabulary				
Iowa Tests of Educational Development: SRA Assessment Survey	9-12	1974	SRA	MMY 20
Reading				
Mathematics				
Language Arts				
Social Studies				
Science				
Metropolitan Achievement Tests (METRO '78)	Kindergarten-12	1978	Psy. Corp.	Fall 78 NCME
Reading Comprehension				
Mathematics				
Language				
Social Studies				
Science				

CONTINUED ON FOLLOWING PAGE

MULTISUBJECT ACHIEVEMENT BATTERIES

Tests and Subscores	Grade Level(s)	Publication Date	Publisher	Reference
National Educational Development Tests	7-10	1974	SRA	MMY 23
Mathematics Usage				
English Usage				
Social Studies Reading				
Natural Sciences Reading				
Word Usage				
Primary Survey Tests	2-3	1973	SF	TIP 27
Reading				
Mathematics				
Language				
Spelling				
Scholastic Testing Service Educational Development Series Scholastic Tests	2-12	1976	STS	MMY 20
Reading				
Mathematics				
English				
Social Studies				
Science				
Solving Everyday Problems				
USA in the World				
Nonverbal Ability				
Verbal Ability				
School Interests				
School Plans				
Career Plans				
Science Research Associates Achievement Series (ACH) Forms 1&2	Kindergarten-12	1978	SRA	NCME Fall 78
Reading				
Mathematics				
Language Arts				
Social Studies				
Science				
Reference Materials				
Applied Skills				
Science Research Associates High School Placement	9	1973	SRA	TIP 31
Reading				
Arithmetic/or Modern Math				
Language Arts				
Social Studies				
Science				
Science Research Associates Norm Referenced/Criterion Referenced Testing Program	4-10	1977	SRA	TCB July 77
Reading				
Mathematics				

CONTINUED ON FOLLOWING PAGE

Stanford Diagnostic Reading Test (SDRT)	1-13	1976	Psy. Corp.	MMY 777
Wisconsin Design for Reading Skill Development	K-6	1972	NCS	MMY 778
Woodcock Reading Mastery Tests (WRMT)	K-12	1973	AGS	MMY 779

MATHEMATICS TESTS

Test	Grade Level(s)	Publication Date	Publisher	Reference
Analysis of Skills: Mathematics (ASK: Mathematics)	1-8	1976	STS	MMY 251
Assessment of Skills in Computation (ASC)	7-9	1978	CTB	NOT Oct 79
Basic Arithmetic Skill Evaluation	1-9	1974	IILC	MMY 303
Diagnosis: An Instructional Aid: Mathematics	1-6	1974	SRA	MMY 263
Diagnostic Mathematics Inventory (DMI) (Revision of the PMI)	1-8	1975	CTB	MMY 264
Diagnostic Screening Test: Math (DSTM)	1-11	1979	SC	NOT Nov 79
ERB Modern Arithmetic Test	5-6	1971	ERB	TIP 718
Fountain Valley Teacher Support System in Mathematics (FVTSS-M)	K-8	1974	Zweig	MMY 270
Individualized Criterion Referenced Testing: Math (ICRTM)	1-8	1977	EDC	MMY 275-6
Individual Pupil Monitoring System-Mathematics (IPMS)	1-8	1973	RS	MMY 274
Keymath Diagnostic Arithmetic Test	K-6	1976	AGS	MMY 305
Mastery: An Evaluation Tool: Mathematics	K-9	1976	SRA	MMY 278
Mathematics: IOX Objectives-Based Tests	K-9	1976	IOX	MMY 279
Minimum Essentials for Modern Math	6-8	1971	Hayes	TIP 638
Objectives-Referenced Bank of Items and Tests: Mathematics (ORBIT:M)	K-Adult	1975	CTB	MMY 287
Stanford Diagnostic Mathematics Test (SDMT)	1-Adult	1976	Psy. Corp.	MMY 292
Steenburgen Quick Math Screening Test	1-6	1978	ATP	NOT Feb 79
Tests of Achievement in Basic Skills: Mathematics (TABS:M)	K-12	1976	EdITS	MMY 293

LIFE SKILLS TESTS

Test	Grade Level(s)	Publication Date	Publisher	Reference
Adult Performance Level Functional Literacy Test	9-Adult	1978	ACT	FLIT Pg. 42
Assessment of Skills in Computation (ASC)	7-9	1978	CTB	NOT Oct 79
Everyday Skills Tests (EDST)	6-12	1975	CTB	MMY 18
IOX Basic Skills Test	9-12	1978	IOX	NCME Spring 1979
NM Consumer Mathematics Test	9-12	1973	NMS	MMY 312
Reading/Everyday Activities in Life (R/EAL)	9-Adult	1972	CAL-P	FLIT Pg. 46
SRA Coping Skills: A Survey plus Activities	7-8 & Adult	1979	SRA	NCME Special Edition 1979

SRA Survival Skills	6-Adult	1976	SRA	TCB Jul 77
STS Educational Development Series: Scholastic Tests	2-12	1976	STS	MMY 27
Senior High Assessment of Reading Performance Forms A, B, C (SHARP)	10-12	1978	CTB	TCB Jul 77 Pg. 11 (Form A) NCME Winter 77 (Form B)
Stories about Real-Life Problems	5-8	—	NIU	NOT May 79
Test of Consumer Competencies	8-12	1976	STS	TCB Jan Pg. 6
Test of Everyday Writing Skills (TEWS)	9-12	1978	CTB	NCME Spring 78
Tests of Performance in Computational Skills (TOPICS)	9-12	1978	CTB	NCME Winter 77
Wisconsin Test of Adult Basic Education	Adult	—	RFD	FLIT Pg. 48

LANGUAGE ARTS TESTS

Test	Grade Level(s)	Publication Date	Publisher	Reference
Analysis of Skills: Language Arts (ASK: Language Arts)	2-8	1976	STS	MMY 41
Diagnostic Screening Test: Language	K-Adult	1977	SC	NOT Oct 79
Language Arts: IOX Objectives-Based Tests	K-6	1974	IOX	MMY 53
Language Arts: Minnesota High School Achievement Examinations	7-12	1970	AGS	TIP 90
Writing Test: McGraw-Hill Basic Skills System	11-12, Adults	1970	MHBC	TIP 125

Publishers' Names, Addresses and Telephone Numbers

This Appendix lists all publishers identified in Appendix A.

ACT The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52240
(319) 356-3711

AGS American Guidance Service, Inc.
Publisher's Bldg.
Circle Pines, MN 55014
(612) 786-4343

ATP Academic Therapy Publications
28 Commercial Blvd.
Novato, CA 94947
(415) 883-3314

AW Addison-Wesley Publishing Co., Inc.
Jacob Way
Reading, MA 01867
(617) 944-3700

BFA BFA Educational Media
2211 Michigan Avenue
P.O. Box 1795
Santa Monica, CA 90406
(213) 829-2901

BMC Bobbs Merrill Co., Inc.
4300 West 62nd Street
Indianapolis, IN 46268
(317) 298-5400

CAI Curriculum Associates, Inc.
5 Esquire Rd.
N. Billerica, MA 01862
(617) 935-8410

CAL-P CAL Press, Inc.
76 Madison Ave.
New York, NY 10016
(212) 685-0892

CARE The Center for Applied Research in Education, Inc.
Rt. 59
West Nyack, NY 10994
(914) 358-8991

Croft Croft Incorporated
4922 Harford Road
Baltimore, MD 21214
(301) 254-5082

CSDE California State Dept. of Education
721 Capitol Mall
Sacramento, CA 95814
(916) 445-4688

CTB CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940
(408) 649-8400

EDC Educational Development Corporation
P.O. Box 45663
Tulsa, OK 74145
(918) 622-4522

EdITS EdITS/Educational and Industrial Testing Service
P.O. Box 7234
San Diego, CA 92107
(714) 222-1666

ERB Educational Records Bureau
Educational Testing Service
Box 619
Princeton, NJ 08540
(609) 921-9000

EPS Educators Publishing Service
75 Moulton St.
Cambridge, MA 02138
(617) 547-6706

HAYES Hayes Educational Laboratory
7040 North Portsmouth Ave.
Portland, OR 97203
(503) 285-3745

IILC Imperial International Learning Corp.
Box 548
Kankakee, IL 60901
(815) 933-7735

IOX Instructional Objectives Exchange
Box 24095
Los Angeles, CA 90024
(213) 474-4531

Jastak Jastak Associates, Inc.
1526 Gilpin Ave.
Wilmington, DE 19806
(302) 652-4990

McGrath McGrath Publishing Co.
P.O. Box 9001
Wilmington, NC 28402
(919) 763-3757

Merrill Charles E. Merrill Publishing Co.
1300 Alum Creek Drive
Columbus, OH 43216
(614) 258-8441

MHBC McGraw Hill Book Co.
1221 Ave. of the Americas
New York, NY 10020
(212) 997-1221

NCS NCS Interpretive Scoring Systems
4401 West 76th St.
Minneapolis, MN 55435
(800) 328-6290

NIU Northern Illinois University
Alan M. Voelker
Curriculum and Instruction
Dekalb, IL 60115
(815) 753-1000

NMS New Mexico State Dept. of Education
Monitor
Education Bldg.
State Capitol
Santa Fe, NM 87501
(505) 827-2429

Psy. Corp. The Psychological Corporation
304 E. 45th Street
New York, NY 10017
(212) 888-3500

RFD Rural Family Development Program
University Extension
University of Wisconsin
P.O. Box 1379
Madison, WI 53701
(608) 262-1234

RS The Riverside Publishing Company
1919 South Highland Avenue
Lombard, IL 60148
(312) 629-9700

SC Stoelting Co.
1350 S. Kostner
Chicago, IL 60623
(312) 522-4500

SRRC Southwest Regional Resource Center
127 South Franklin
Juneau, AK 99801
(907) 586-6806

SF Scott Foresman & Co.
1900 East Lake Ave.
Glenview, IL 60025
(312) 729-3000

SOI SOI Institute
214 Main St.
El Segundo, CA 90245
(213) 322-5995

SRA Science Research Associates, Inc.
155 N. Wacker Dr.
Chicago, IL 60606
(800) 621-0664

STS Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, IL 60106
(312) 766-7150

TCP Teachers College Press
Teachers College
525 West 120th St.
New York, NY 10027
(212) 678-3929

USDL United States Dept. of Labor
Bureau of Labor Statistics
1515 Broadway
New York, NY 10036
(212) 399-5405

Winch B. L. Winch and Associates
45 Hitching Post Dr.
Rolling Hills Estates, CA 90274
(213) 547-1240

Zweig Richard L. Zweig Associates, Inc.
20800 Beach Blvd.
Huntington Beach, CA 92648
(714) 536-8877

Reference Materials Describing and Reviewing the Tests

Mental Measurements Yearbook (MMY)

Of all sources cited, *The Eighth Mental Measurement Yearbook* provided the most comprehensive information on tests. To best utilize the source, read the introductory section, "How to Use This Yearbook." Tests are indexed in the yearbook by test number. That number is provided in the previous test lists as the follow-up reference number. However, the number of any test can also be located via the yearbook title or subject indices. Information provided about each test includes:

Title
Description of the groups for which the test is intended
Date of copyright or publication
Acronym
Part scores
Individual or group test
Forms, parts, and levels
Pages
Machine-scorable answer sheets
Costs
Scoring and reporting services
Time
Author
Publisher
Foreign adaptation
Sublisting
Cross references

Additional references to published articles, books and unpublished theses on the construction, validity, use and limitations of each test are reported as part of each test entry. Original reviews of each test by independent measurement experts are provided.

Tests in Print (TIP)

The companion volume to the *Yearbook* is *Tests in Print II*. It provides the reader with similar but much less detailed information on tests. Again, a section entitled "How to Use This Book" is provided. *Tests in Print II* presents a bibliography of all known tests published for English-speaking subjects and an index to all tests published in previous editions of the *Mental Measurement Yearbook*. TIP II provides the following information:

Title
Test population
Copyright date
Acronym
Special comments
Part scores
Author
Publisher
Foreign adaptations
Cross references within TIP II
Sublistings

NCME Measurement News (NCME)

The "NCME Measurement News," the official newsletter of the National Council on Measurement in Education, provides a brief description of recently published tests. Information includes publisher, copyright, subject matter, levels, grade, interpreting manuals and costs. It is suggested that individuals desiring additional information contact the publisher directly using the addresses provided. Tests appearing in the newsletter in the newsletter do not represent endorsement by the NCME or its staff.

News on Tests (NOT)

ETS "News on Tests" and its predecessor, the "Test Collection Bulletin," provide descriptions similar to the NCME newsletter. The test title, author, publisher and address copyright and grade level are included, with a brief statement of content and levels. Information provided is descriptive rather than evaluative, and Educational Testing Service "News on Tests" also includes announcements of new publications relating to testing, conferences, and available test bibliographies.

Tests of Functional Literacy (FLIT)

The review of currently available *Tests of Functional Adult Literacy* provides information on the characteristics and quality of a series of standardized, criterion referenced, and informal tests of literacy. Included in the descriptive profiles of tests is information on publisher, content and skill coverage, availability of alternate forms, administration procedures, materials needed, scoring procedures, interpretation procedures, validity and reliability.

The reader is urged to take advantage of these informational documents and to contact test publishers for complete information on available tests.

Major U.S. Publishers of Standardized Tests from the Test Collection, Educational Testing Service

The following publishers are listed in this collection which were not listed in the collection supplied by the Northwest Regional Educational Laboratory.

Bureau of Educational
Measurements
Kansas State Teachers College
Emporia, KS 66801
316-343-1200

Bureau of Educational
Research & Service
C-20 East Hall
The University of Iowa
Iowa City, IA 52240
319-353-2823

Committee on Diagnostic
Reading Tests, Inc.
Mountain Home, NC 28758
704-693-5223

Consulting Psychologists Press
577 College Avenue
Palo Alto, CA 94306
415-326-4448

Educational Testing Service
Princeton, NJ 08541
609-921-9000

Western Office:
1947 Center Street
Berkeley, CA 94704
415-849-0950

Follett Publishing Co.
A Division of Follett Corp.
Department DM
1010 West Washington Blvd.
Chicago, IL 60607
312-666-5855

Ginn and Company
P. O. Box 2649
1250 Fairwood Avenue
Columbus, OH 43216
614-253-8661

Grune and Stratton, Inc.
111 Fifth Avenue
New York, NY 10003
212-741-6800

Guidance Testing Associates
of St. Mary's University
1 Camino Santa Maria
San Antonio, TX 78284
512-436-3304

Institute for Personality and
Ability Testing (IPAT)
1602 Coronado Drive
Champaign, IL 61822
217-352-4739

Martin M. Bruce, Publishers
340 Oxford Road
New Rochelle, NY 10804
914-235-4450

Priority Innovations, Inc.
P. O. Box 792
Skokie, IL 60076
312-729-1434

Psychological Research Services
Case Western Reserve University
1695 Magnolia Drive
Cleveland, OH 44106
216-368-3536

Psychological Test Specialists
Box 1441
Missoula, MT 59801

Psychologists and Educators, Inc.
Suite 212
211 West State Street
Jacksonville, IL 62650
217-243-2135

Psychometric Affiliates
Box 3167
Munster, IN 46321
219-836-1661

Richardson, Bellows, Henry
and Co., Inc.
1140 Connecticut Ave.
Washington, DC 20036
202-659-3755

Sheridan Psychological
Services, Inc.
P. O. Box 6101
Orange, CA 92667
714-639-2595

University Bookstore
Purdue University
360 State Street
West Lafayette, IN 47906

Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, CA 90025
213-478-2061

RESEARCH
REPORT

WORKBOOK

PLAIN TALK ABOUT STANDARDIZED TESTS

BY

JAMES G. WARD
DIRECTOR OF RESEARCH

JEWELL C. GOULD
ASSISTANT DIRECTOR OF RESEARCH

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

TM 810 625 (2 of 2)

[illegible]

A REPORT OF THE
RESEARCH DEPARTMENT OF THE
AMERICAN FEDERATION OF TEACHERS, AFL-CIO
OCTOBER, 1980

EXERCISES

Problem #1

Using as much of the data about the district and tests available, select the test that best suits your district's needs. Your goals for testing should be determined and the only limitation is that you are under a mandate to offer general evaluation data to the schools and public through the grades. The evaluation instrument selected must be supported by whatever facts you can draw from the information. Your process of selection and other considerations should be noted.

Problem #2

You have selected a test and the scores are reported for the district, schools, for a class, and for some students. Groups one and two will evaluate the scores for the district and prepare a statement to present to the district. They must be prepared to answer any questions regarding their analysis and should prepare a short presentation which will be given at (1) either the district board of education meeting or (2) a press conference called to announce the results of the testing program.

Groups three and four will analyze the class data. Group three should be prepared to discuss the use of the data in making generalizations about the class performance, strengths and weaknesses of pupils, and recommendations for individual students.

Group four will evaluate the test scores in light of proposed curriculum changes or additions that may be necessary. They will meet with a parent council to discuss the necessary changes in light of problems identified with subject matter areas.

Group five and other groups will review the scores of students listed on the school's score sheet. They will prepare comments for parents of the children and present four of the eight in the final review. If necessary they must defend the scores and the test.

HARMON, U.S.A.

Population 1975
14285

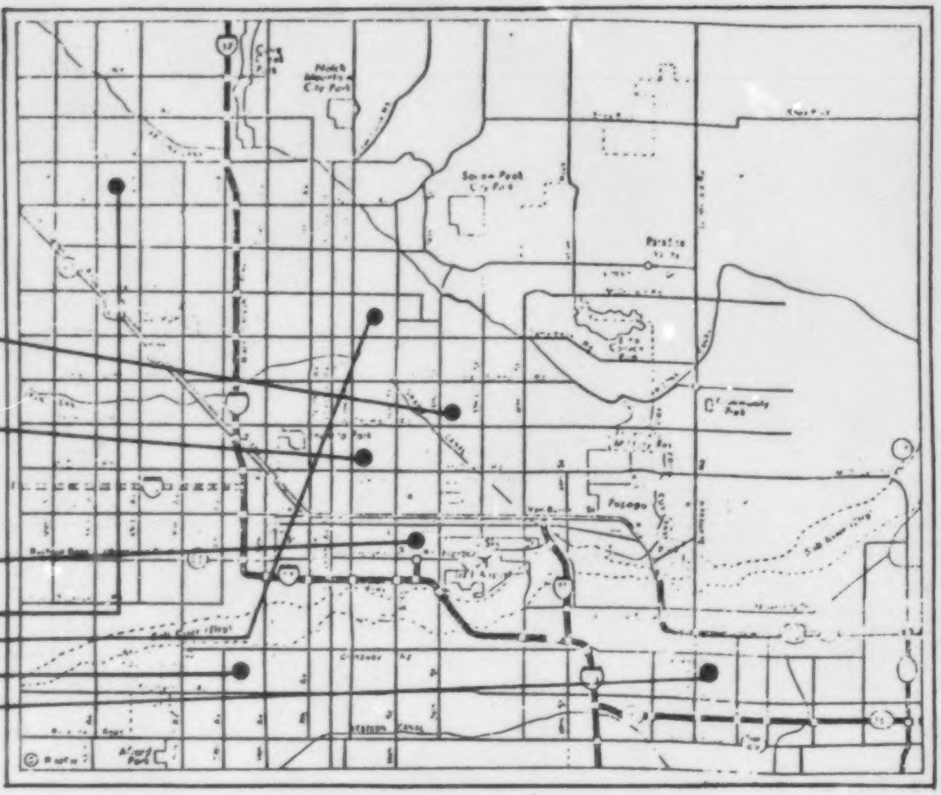
Number of Students
5480

Number of Schools
7

High School, Dewey
1515 students; 55
teachers; 19.5:1 PTR

Junior High, Kennedy
1165 students; 64
teachers; 18:1 PTR

Elementary Schools -
190 teachers;
Duff - 600
Chiddix - 656
North - 700
Brett - 625
Rose - 650
PTR: 17:1



Harmon, U.S.A., is a town with a high school, junior high, and five elementary schools. The enrollment in recent years has been declining slightly, but in most respects it is an average small district near an urban center. It has a manufacturing plant, assorted medium and small businesses offering services to the residents, a state college, and is served by major highways, railroads, and boasts one airport. Sky Harbor International Airport.

The schools have grown up around the city. The Junior High occupies the former high school building. The airport was built adjacent to the Duff Elementary School on land acquired by the district and later sold to the city when enrollment declines offset the need for further construction. As a consequence the development of housing has been pressing to the east part of town and to the north. Elementary enrollment has been increasing slightly and changing in composition of the student body in recent years.

The spendable family income of $10,360 is slightly below the national average of $10.504. While the national population's education level indicates that 52 percent graduate from high school, in Harmon 54 percent have completed high school. The ethnic breakdown for the community is 71 percent white, 19 percent black, 6 percent hispanic and 4 percent other. Recently a sizable group of refugees have relocated in the city and their children have enrolled in public schools.

**HARMON
SCHOOL
DISTRICT**
11 DUPONT CIRCLE, N.W., WASHINGTON, D.C. 20036

MEMORANDUM

To: Staff

From: Test Evaluation Committee

Re: Technical Data Comparison

Date: December 3, 1979

The subcommittee on test comparison met and reviewed a number of tests over the last three months. We have determined that one battery of tests will be best for our purposes as outlined by the committee on evaluation in their memo last spring (attached). In keeping with the direction of the committee, however, to present three choices for final comparison, we are reviewing in this memo the Standardized Achievement Series Assessment Survey (SASAS), the Urban School Estimate (USE), and the Criterion Ability Check of Potential (CRACPOT). Copies of the various tests are available for your consideration. In addition we have compared some major points on each of the tests and summarized the information.

We have taken our information from the test manuals, from Buros Mental Measurement Yearbooks, and from publications in which the tests were reviewed. Where applicable, we have included comments from those documents.

TEC/ag
opeiu#2aflcio

COMPARISON OF TESTS
HARMON SCHOOLS TEST EVALUATION SUBCOMMITTEE

Item	SASAS	USE	CRACPOT
Reliability			
Test-Retest	.87	.85	Not available
Equivalent forms	Not available	Not available	Not available
Split Half (RII)	.91	Not available	.86
KR 20	.75	.67	.86
KR 21	Not available	Not available	Not available
Standard Error of Measurement	2.4	2.9	3.1
Validity			
Content	Acceptable match to district - 2nd closest of those reviewed	Fails to match our language objectives, Math	Closest to our curriculum & processes
Criterion Related			
Predictive *	.85 correlation with future test scores in sample test of 100 fall administration	.73 correlation with spring scores on fall administration to sample of 100	.60 correlation with spring scores after fall administration to sample Of 100
Concurrent	Predicted stanine ranges in 95% of the cases on Reading test	Predicted stanine ranges in 80% of the cases	Stanine ranges not available except by dividing percentile ranks
Construct	The committee reviewed the test manuals and other reviews and are satisfied that SASAS and USE will be compatible with our district philosophy. CRACPOT was too oriented toward single score progress identifications		
Face	Test appeared to be taken seriously in sample; was easy to read, mark, and acceptable	Students were somewhat confused by presentation of math computations; oral response portions were subject to variations with administrators of tests; some portions acceptable but answer sheet difficult to follow for students resulting in some obvious mismarking of choices.	Use of pic-to-grams failed to give students confidence in test. Test booklet was printed in light green and purple, making it difficult to read. Not acceptable.

---

* Sample test was administered to 100 students in fall and test currently used administered in spring of following year. The reading level test was administered at the same time in the fall and stanine ranges were compared.

COMPARISON OF TESTS
HARMON SCHOOLS TEST EVALUATION SUBCOMMITTEE

Item	SASAS	USE	CRACPOT
Norm Data			
Student sample size	275,000	200,000	1,950
Standardization Month, year	April, October 1977	May, September 1975	January 1979
States or Regions	Proportionate allocation as to U.S. Census Report	12 Major urban cities in U.S.	50% North Central U.S. 40% Western U.S. 10% Eastern U.S.
Community Size	35% Urban, 50% Small Town or Suburban, 15% Rural	100% Cities 500,000 or more	60% Small town, 20% Rural, 20% Urban
Sex	48.6% Male 51.4% Female	48% Male 52% Female	54% Male 46% Female
Age	5% each ages 5-7; 16-17; 75% ages 8-15 evenly divided	4% age 5; 8% all ages 6-17	Ages 5-12 evenly divided
Race	White 70%; Black 20; Other 10%	White 37%; Black 54%; Other 9%	White 80%; Black 18%; Other 2%
Other Data			
Scores Reported	Percentile Ranks, Scaled Scores, Raw Scores	Raw Scores, Percentile Ranks, Stanines	Raw Scores, Percentile Ranks, (P) Values
Cost (per pupil includes all manuals)	$1.55 + $.40 for ability test recommended	$.75	$1.10
Time to Administer	3 hrs. plus 50 mins. for ability test	1 hr. 45 mins.	40 mins. plus scoring
Scoring Services	$.35 each machine scoring Various services	$.10 - $.60 depending on machine scoring chosen	hand scoring only

Standardized Achievement Series Assessment Survey (SASAS)

Results of the SASAS Assessment Survey, administered in October 1979 to Harmon public school students in grades 4, 6, 8, and 11, are reported on the following pages. The scores reflect the average performance in regular programs; those obtained by children in self-contained programs (such as gifted/talented and learning disabled) are not included in school averages.

The SAS Assessment Survey is a norm-referenced ability/achievement test published by Simpson, Sampson and Belinski Associates. It is designed to sample students' achievement of the concepts and skills common to most school programs. A brief description of the content of the subtests follows.

The Standardized Estimate measures general educational ability based upon those factors most closely associated with academic performance, i.e., measures of verbal, number, and reasoning abilities. It is designed to assess the student's present academic aptitude.

The Reading Comprehension test measures the ability to understand central themes and main ideas, draw logical conclusions, and retain significant details. The selections represent several subject areas: fiction, biography, science, and social studies.

The Reading Vocabulary test measures recognition of synonyms for short phrases and knowledge of words as they appear in written context.

The Language Arts Usage test measures knowledge of basic elements required for correct and effective writing. Included are capitalization, punctuation, and sentence and paragraph structure; use of modifiers, nouns, verbs, and pronouns; and diction.

The Language Arts Spelling test measures recognition of misspelled words.

The Mathematics Concepts test measures understanding of basic numeration and mathematical operations plus knowledge and application of concepts in measurement, geometry, and problem solving.

The Mathematics Computation test measures ability to handle computational operations involving addition, substraction, multiplication, and division of whole numbers, whole number groups, fractions, decimals, and percents.

The Mathematics test (grade 11) presents exercises involving practical, realistic situations as well as more abstract exercises involving number systems and more sophisticated mathematical concepts.

The Social Studies test measures knowledge and appropriate application of concepts in geography, history, economics, sociology, anthropology and political science plus the ability to use written and illustrated materials.

The Science test measures knowledge and appropriate application of concepts in biology, matter and energy, earth and space, and experimentation plus the ability to use written and illustrative materials.

The Uses of Sources test measures ability to use basic sources of information, such as tables of contents, indexes, dictionaries, reference books, library catalogues, maps, charts, and graphs.

Urban School Estimate (USE)

The USE test was investigated by the test committee and a sample of students tested in the process. We found it to be a competitive test in many respects for the regular classroom population. It is a norm-referenced achievement test published by Center Evaluation and Measurement Systems, Inc. of Metroville, Ohio. It is designed to sample student achievement throughout grades K-12. It is intended primarily for the urban market and takes into account some of the specialized programs found in these districts. Often such programs are progress-linked as opposed to traditional grade advancement progress measures that require specific curricular content at each grade level. The USE test is designed to take advantage of this kind of curricular system and allows the test user to reference to national norms as well as establish the level of achievement of the student. Sensitive measures of the student's ability and achievement are possible in districts where the curriculum closely matches the test's content.

Prior to the development of the test items, extensive survey and examination of major urban districts' curricula was made. While each district has its own characteristics, the authors believe that a majority of districts who have been in the move back to basics will benefit by use of the USE test in their district. Math and language objectives in these areas are geared to the recently published series of texts by Bates-Universe Press which have become the standard for basic education programs.

The Reading test covers vocabulary, comprehension, and provides a total score for the section. Grade level may be assigned, but the test recommends that levels in the Bates-Universe series be utilized for inferences about level. The vocabulary has been reviewed to eliminate potential bias and ambiguity that many tests include due to their orientation to standard English without regard to bilingual students or students who are familiar with non-standard English.

The Mathematics section also follows the Bates-Universe series and levels can be established independent of reading scores. The section on series and relations of numbers fits quite well with more standard conceptual and computational sections of tests.

Writing skills emphasizes spelling and proofreading, common errors of students, and relies on the vocabulary and spelling words in the Bates-Universe series for 80 percent of the words. Distractors in the proofreading portion of the test are uniformly difficult and speed is stressed.

Study skills with some map reference, dictionary usage, and reference material location are tested in this section.

Criterion Ability Check of Potential (CRACPOT)

CRACPOT, a new test published in 1979 is perhaps the test we all have been waiting for. Because this test is not referenced to any norm group, but rather to criteria, no more can students be burdened with the knowledge that they are below average or in the failure mode. Simply test, test, and test again until the student gets it right. By use of the hand scoring system and the relatively short administration time for the test, the teacher can pinpoint precisely what the student has learned and what should be studied.

Critics of district test scores will have to adjust their sights for some time to come when you begin reporting the numbers of students who are able to successfully pass the test. Progressive difficulty levels enable you to test the student for any selected criterion level and students can actually expect to pass provided the level is suited to their rate of learning.

Alternate score reporting formats are available to suit purposes of the district. The Standard Operating Scores format is a scaled score system that will allow districts to report specific large numbers to the public and provide teachers with percent correct information. Hand scoring encourages the teachers to get in touch with their students and saves a considerable amount of money over machine scored services. The turnaround time for scoring these tests depends upon how dedicated the teachers are or upon the amount of leadership pressure exerted by district officials, so anything is possible.

The test was standardized throughout the country and in the opinion of the authors, most districts will be able to match their curriculum objectives to the various test items. Quick, inexpensive, and sensitive to the problems of educators today, the CRACPOT will get the public off your back once and for all.

Standardized Achievement Series Assessment Survey (SASAS)
National Percentile Equivalents for Mean Scores
(National average 50th percentile)

				READING			LANGUAGE ARTS			MATH					
School	Grade	Ability	Composite	Comprehension	Vocabulary	Total	Useage	Spelling	Total	Concepts	Computation	Total	Social Studies	Science	Use of Sources
Elementary															
Duff	4	44	43	48	42	43	50	47	47	52	54	54	45	47	45
	6	52	64	50	52	49	61	61	59	58	69	71	48	54	64
Chiddix	4	85	87	84	87	86	86	78	85	80	81	81	87	85	87
	6	85	84	81	85	84	83	80	83	87	72	80	80	76	88
North	4	62	59	58	60	61	69	62	62	62	43	54	63	59	67
	6	55	62	61	61	61	61	61	59	66	49	57	45	42	48
Brett	4	49	53	58	57	57	58	57	56	52	49	51	56	51	58
	6	49	59	53	49	51	59	61	58	58	57	57	45	54	62
Rose	4	44	35	37	37	35	44	37	37	40	32	35	36	38	40
	6	37	39	40	31	34	37	42	38	36	49	41	34	31	43
Junior High															
Kennedy	8	58	61	56	53	54	55	54	56	62	63	64	50	49	61
Senior High															
Dewey	11	59	65	55	62	62	60	48	53	68	64	66	59	55	54
Student															
Chop, M.	4	85	86	78	80	81	83	74	79	85	87	89	79	97	83
Cartwright, C.	4	49	50	58	50	53	52	62	54	43	52	45	45	51	49
Meyers, H.	6	83	80	80	78	81	69	74	70	83	77	83	72	73	70
Turner, J.L.	6	52	54	58	52	50	59	55	57	49	45	48	55	51	62
Southern, B.	8	80	84	84	81	84	78	75	78	87	84	89	82	81	85
Duncan, D.	8	70	69	69	66	73	62	71	62	71	68	71	72	67	71
Martinez, R.	11	55	75	61	61	62	74	61	68	81	89	86	61	63	73
Ku, A.L.	11	53	43	38	29	34	37	24	28	40	58	52	43	49	45

Standardized Achievement Series Assessment Survey (SASAS)

National Percentile Equivalents for Mean Scores

(National average 50th percentile)

Students Grade 4	Ability	Composite	READING			LANGUAGE ARTS			MATH			Social Studies	Science	Use of Sources
			Comprehension	Vocabulary	Total	Useage	Spelling	Total	Concepts	Computation	Total			
R.F.	44	43	48	42	43	50	47	47	52	54	54	45	47	45
T.C.	52	64	50	52	49	61	61	59	58	69	71	48	54	64
J.B.	62	59	58	60	61	69	62	62	62	43	54	63	59	67
R.N.	85	84	81	85	84	83	80	83	87	72	80	80	76	88
C.N.	85	87	84	87	86	86	78	85	80	81	81	87	85	87
B.P.	55	62	61	61	61	61	61	59	66	49	57	45	42	48
K.K.	49	53	58	57	57	58	57	56	52	49	51	56	51	58
S.T.	44	35	37	37	35	44	37	37	40	32	35	36	38	40
E.W.	44	35	37	37	35	44	37	37	40	32	35	36	38	40
E.G.	37	39	40	31	34	37	42	38	36	49	41	34	31	43
J.D.	23	24	21	18	19	28	12	11	13	29	18	26	19	15
A.L.	58	61	56	53	54	55	54	56	62	63	64	50	49	61
J.F.	44	43	48	42	43	50	47	47	52	54	54	45	47	45
M.G.	52	64	50	52	49	61	61	59	58	69	71	48	54	64
J.W.	59	65	55	62	62	60	48	53	68	64	66	59	55	54
H.J.	67	64	63	65	66	74	67	67	67	48	62	68	50	72
E.P.	55	75	61	61	62	74	61	68	81	89	86	61	63	73
W.S.	53	43	38	29	34	37	24	28	40	58	52	43	49	45
M.D.	83	80	80	78	81	69	74	70	83	77	83	72	73	70
M.O.	52	54	58	52	50	59	55	57	49	45	48	55	51	62
D.P.	49	50	58	50	53	52	62	54	43	52	45	45	51	49
G.M.	70	69	69	66	73	62	71	62	71	68	71	72	67	71
M.S.	85	86	78	80	81	83	74	79	85	87	89	79	97	83
M.T.	70	69	69	66	73	62	71	62	71	68	71	72	67	71

Stanine Ranges for SASAS and Ability Test

Percentile Rank	Stanine
99 - 96	9
94 - 89	8
88 - 77	7
76 - 60	6
58 - 40	5
38 - 23	4
22 - 11	3
10 - 4	2
2 - 1	1